BRASFIELD, JON, Ph.D. A Comparison of Classification Issues across Teacher Effectiveness Measures. (2014)
Directed by Dr. Terry Ackerman. 140 pp.

In an educational landscape where teacher evaluation methods are increasingly discussed and scrutinized in research offices, legislatures, and school buildings, the differences in policy and instrumentation among states and school districts can paint a confusing picture of these varying methods and their impacts. To help assess the picture in North Carolina, this project examined teacher effectiveness data on 147 teachers from 16 schools in a large urban school district. Three measures of teacher effectiveness (a value-added measure of student growth, a third-party observation score, and state-mandated principal evaluations) were examined with a particular focus on how teachers were classified via the different methods.

The first research question examined the similarities and differences in classification across the measures. Correlational and cross-tabular results suggested that the value-added measure and the third-party observational measure were not independent. It was also found that principal ratings do little to differentiate between teachers.

The second question examined the relationships of the variables that determine teacher pay with the effectiveness measures. It was found that no substantive relationships exist, suggesting that the teacher evaluation measures in this sample do not exhibit bias based on experience, advanced degrees, or National Board status.

The third question examined relationships between overall school effectiveness measures and school demographic variables. The third-party observation measure tended

to be moderately associated with school-level demographics, a finding which may call into question the ability to compare teachers across schools.

A COMPARISON OF CLASSIFICATION ISSUES ACROSS TEACHER EFFECTIVENESS MEASURES

by

Jon Brasfield

A Dissertation Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2014

Approved by

_____
Committee Chair

To Amanda, and everyone else who knew I could and would do it.

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair_____

Committee Members_____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to acknowledge the guidance of my advisor and committee chair, Dr. Terry Ackerman, and the other members of my committee, Dr. Kimberley Kappler Hewitt, Dr. Judy Penny, and Dr. John Willse. Thank you for your thoughts, input, and flexibility in this process. I would also like to thank Dr. Marty Ward and the Winston-Salem/Forsyth County school district for their cooperation and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Today, more data on more aspects of education are available than ever before. The ability of states, districts, and educational research organizations to capture and link information on students and teachers in a longitudinal manner has far-reaching implications for the future of educational practice in the United States. One area in which this wealth of data has recently received attention both within education policy arenas and in mainstream society is the field of teacher effectiveness evaluation (e.g. Anderson, 2013; Banchero & Kesmodo, 2011; Ripley, 2010).

Partially due to the amount of information available and partially due to federal grant programs such as Race to the Top (RttT) and the requirement to use student growth data in teacher evaluation systems in order to receive a waiver from No Child Left Behind requirements, states and school districts have been revising their teacher evaluation systems at a rapid pace. In the period from 2009-2012, 36 states and the District of Columbia made substantive changes to their policies regarding teacher evaluation, including guidelines on tenure, retention, dismissal, and in some cases, performance pay (National Council on Teacher Quality, 2013). Though many policy changes have been instituted, there remains much diversity in the policies themselves, as well as in the research regarding the best aspects of a quality teacher evaluation system. Traditionally, teacher evaluation has been conducted as a primarily building-level

exercise, with a principal or other administrator issuing a rating of proficient or nonproficient on a teacher, and these ratings having some nominal effect on decisions regarding tenure status.

In recent years, considerable research has been done on the efficacy of this particular method of teacher evaluation, as well as on emerging methods of teacher evaluation such as value-added modeling (VAM), standards-based observation, and student surveys, among others. In 2011, the Bill and Melinda Gates Foundation completed a massive study on this topic in an attempt to determine the best methods of measuring teacher quality. Among their findings was that standards-based observations, student survey data, and value-added modeling of teacher effects on student learning could be valuable indicators of quality. Policy changes across states have been inspired by this research, with 39 states now requiring that classroom observations of teachers be incorporated into teachers' evaluations and 30 states requiring that some measure of student achievement be a significant or the most significant aspect of a teacher's evaluation (Collins & Amrein-Beardsley, 2013). According to a nationwide survey of educational effectiveness policy, 40 states and the District of Columbia currently are either using or developing some sort of growth measure or VAM (Collins & Amrein-Beardsley, 2013). In the past four years, the number of states requiring that student achievement play a role in teacher tenure decisions grew from zero to nine (NCTQ, 2013). In addition, where teachers previously were evaluated irregularly, 23 states now require that every teacher undergo annual evaluation.

With so much research being done on the effectiveness of various methods of measuring teacher effectiveness, opportunities arise for examination of the purpose of teacher evaluation. Papay (2012) proposed that evaluation systems serve two potential purposes: first, to identify high- and low-quality teachers for purposes of tenure, dismissal, or merit pay; and second, to serve as formative assessment for a teacher's practice, providing him or her with feedback crucial to the professional growth process. For a teacher evaluation system to serve either of these purposes, the conclusions reached by the use of these instruments must be meaningful and provide some diagnostic or prescriptive information. Different instruments used to sort teachers into performance-based groups would ideally sort teachers into the same groups, regardless of the influence of outside factors. In this project, three different teacher effectiveness measures from a set of elementary and middle schools in a large, urban school district in North Carolina were examined. The data were analyzed to assess how various teacher- and school-level variables affect measurements under various methods, as well as to determine if certain categories of teachers might be scored dissimilarly on various methods.

Teacher Evaluation in North Carolina

In North Carolina (at the time of data collection), teachers are subject to annual evaluation using the North Carolina Teacher Evaluation System (NC TES), a standards-based instrument designed to be completed by administrators (see Chapter 3). The system features five standards – one each regarding leadership, establishing a respectful environment for all students, content knowledge, facilitation of learning, and reflecting

on their practice. Veteran teachers are evaluated annually on two standards of the NC TES instrument (leadership and facilitation of learning), and beginning teachers and teachers in review years are evaluated on all five standards. In their 2012 "State of the States" analysis of teacher evaluation practices nationwide, the National Council on Teacher Quality assigned North Carolina a grade of "C-" for its ability to identify quality teachers – a grade that, while low, was in the top half of states. The grade was based on NCTQ's interpretation of North Carolina's policies as requiring evidence of student learning, but it lacked the inclusion of student growth as a "significant" factor in determining teacher tenure. As of 2012 in North Carolina, teacher tenure was granted almost automatically after four years of service (NCTQ, 2013). Legislative action in North Carolina has since begun phasing out teacher tenure, and the status will be fully eliminated in 2018. In its recommendations, the NCTQ suggested that states use multiple measures of student learning to measure teacher effectiveness and require classroom observations that focus on and document the effectiveness of instruction, stating, "well-designed and executed observations provide the clearest opportunity to give teachers actionable feedback on the strengths and weaknesses of their instructional practice" (NCTQ, 2013, p.8).

Whereas the NC TES does require evidence of student learning and does require classroom observations, it does not track the fidelity of observations through records of inter-rater reliability. The data set used in this project involved a second set of standards-based observations conducted by full-time, highly trained observers who exhibited high inter-rater reliability – one trait not present in the literature on the NC TES.

In addition to the NC TES, North Carolina teachers are also measured by a VAM, SAS Institute Inc.'s Education Value-Added Assessment System (EVAAS®). Though much controversy exists around the use of VAMs (see Chapter 2), the fact that North Carolina currently (and for the foreseeable future) uses EVAAS® for teacher measurement warrants its inclusion in this study.

The subset of schools used in this study is unique in that teachers are evaluated by three measures of teacher effectiveness instead of the typical two. North Carolina publishes aggregate data on NC TES outcomes (http://apps.schools.nc.gov/pls/apex/f?p=155:1), and statewide EVAAS® trends are available to educators in the state, but the addition of a highly-focused standards-based observation system to these schools allows an opportunity to observe the differences in levels of classification between the two existing methods of teacher measurement in the state side-by-side with another method that has shown promise in research. Differences in teacher classification shown among the instruments and methods help contribute to the conversation about the best methods of teacher effectiveness measurement and potentially examine the effects of introducing a standards-based observation component performed by highly trained observers into the NC TES.

As of 2012-13 in North Carolina, teachers were eligible for tenure based on the results of the NC TES evaluations. As previously mentioned, teacher tenure status has recently been removed by the North Carolina Legislature and will be completely eliminated by 2018 under the current model. Teachers who scored below recommended

standards were placed in a probationary category, where termination may occur if improvement is not shown. Teacher pay in North Carolina is not related to the evaluation system, and instead is determined by a teacher's experience, highest degree earned, and National Board of Professional Teaching Standards certification status. For this reason, measurement methods were analyzed in the context of both classification and pay variables. This research was approached as a case study and is only intended to be descriptive of the sample used. The final chapter will discuss implications and directions for further research.

To fully examine the relationships between the methods of teacher measurement used in this sample and the school- and teacher-level variables of interest, the three following research questions were proposed:

RQ1: Do the three methods of teacher effectiveness measurement (TEM) classify teachers in substantively different ways?

This question focused on the distributions created when teachers are classified solely on one of the three methods (NC TES, EVAAS, or STAR3 observation score). Analysis of these distributions allowed examination of the relationships between the methods. Statistical tests were conducted to determine if significant differences existed in the classification of individual teachers using different methods. Of particular interest were the relationships (or lack thereof) in teacher classification via the NC TES and classification via the separate standards-based observation rubric. Differences would suggest that two measures of classroom techniques are measuring different constructs in

practice (or that one or both are unreliable or invalid), which ultimately may suggest that inclusion of a separate measure of classroom technique may be desirable.

RQ2: Are teachers with different degrees of experience, advanced degrees, and National Board Certification - the current methods of determining teacher pay - classified differently using different methods of TEM?

This question examined the distributions created by measuring teachers by each of the three measurement methods and compared them with respect to the variables that currently affect teacher pay in North Carolina. As in RQ 1, the individual teachers were the subjects of interest in this question. If all the instruments used in the measurement system were valid and reliable, and they were shown to classify teachers differently at different levels of the teacher-level variables, there may be implications for determining the fairness of the current teacher pay system in North Carolina.

RQ3: Which, if any, school-level variables have an impact on measurement of teacher quality?

This question focused on the school averages for each of the measurements. Analyses used to answer this question focused on attempting to determine if differences existed in any of the three measurements in terms of sensitivity to school-level factors such as poverty and native English speaking percentage**.** As in RQ1, evidence of differential scoring in this sample would suggest that further research should be conducted concerning the potential bias in the instruments used in evaluating teachers and schools in North Carolina.

Limitations

This study, as mentioned previously, used a highly focused sample from low-performing schools in one urban district in North Carolina. It also only used data from teachers of grades 4-8. This limits the generalizability of conclusions from the analyses. Results from this study can be used to inform directions for further research into the impact of different teacher effectiveness measures used in North Carolina and throughout the United States. In addition, an ideal sample would involve the use of the same instrument by both principals and third-party observers, so that observer-effect and instrument-effect would not be confounded. In this sample, one can only compare the impact of the NC TES with the third-party observer system as it is currently designed.

This study is written in five chapters. The first three examine the purpose, context, data, and procedures of the study. The fourth presents the analysis of the data. The fifth chapter examines conclusions gleaned from the analysis of the data. The answers to the three research questions posed above are addressed and directions for future research are also discussed in Chapter 5.

CHAPTER II

REVIEW OF THE LITERATURE

Teacher Evaluation Systems

Evaluation by principals. Traditionally, teachers have been evaluated by a state- or district-mandated process that involves some degree of judgment based on observable behaviors, both in and outside the classroom. Teachers who score high on these evaluations, in general, tend to also score high on other measures of specific teacher effectiveness, such as classroom management and having better relationships with their students (Stronge, Ward, & Grant, 2011).

A significant amount of research has been conducted on these evaluative processes, and many of these studies have found that principals are the primarily responsible parties for conducting the evaluations (e.g., Zimmerman & Deckert-Pelton, 2003). Studies suggest further that principals are tasked with carrying out teacher evaluations without being well-trained in the process and with an inability to devote sufficient time to fidelity (Toch, 2008). Even when fidelity is attained, the tools used for evaluation often don't focus directly on instructional quality (Toch, 2008).

Despite these challenges, "Principals' behaviors, expectations, and perceptions help build the climate of a school and these data suggest that teachers are looking to their principals to be leaders in this critical domain of assessment and evaluation" (Zimmerman & Deckert-Pelton, 2003, p. 35). The knowledge that the building

administrators are the parties responsible for determining teacher effectiveness contributes to a sense of alertness and attention to quality by teachers, when the evaluations are perceived as useful (Noakes, 2008).

However, many teachers do not find these evaluations to be a positive experience. In one study of 79 teachers in five Florida school districts, 38% of teachers noted that their principals conducted their evaluations in an inconsistent manner; they found the ratings subjective and did not feel that they had ample opportunity to appeal (Zimmerman & Deckert-Pelton, 2003). This same study found that the perception of inconsistency existed both within a school and between schools in a district, and that this inconsistency is a demotivator for experienced teachers.

An additional interview-based study found that elementary principals tend to lack follow-through on evaluations (Arrington, 2010). The author performed a series of four focus group interviews with 6-10 teachers each in Georgia, and concluded that, although reactions to the usefulness of evaluations by their principals were mixed, follow-up was minimal and evaluations were disconnected from a teacher's actual performance.

McConney, Ayres, Hansen, and Cuthbertson (2003) came to similar conclusions regarding the value of such systems. The authors analyzed survey results from the Baltimore, Maryland, public schools district (N = 6,096 teachers) to determine satisfaction with their evaluation system. A large majority of teachers in the system rated "good" to "excellent" the timeliness, dignity, and respect with which their evaluations were conducted, although only around half of teachers rated the accuracy of the evaluation system as "good" or "excellent"). In the same study, only 3 to 4 out of every

10 teachers saw evaluation policies as fair and ethical, applied evenly, oriented toward teacher improvement, and providing for due process. In contrast, McConney et al. found that 7 to 8 of every 10 principals believe that these standards are met to a good or great degree. It appears, then, that a disconnect exists between principals and teachers in their perceptions of the accuracy of teacher evaluation systems.

One potential explanation for teachers' resistance to a principal-led evaluation system is that teaching has traditionally been a solitary activity. Teachers are not socialized to having external observations of their work and are often uncomfortable with the observation process (Van Tassel-Baska, Quek, & Feng, 2006). This phenomenon was clarified in a qualitative study which found that teachers conveyed a "deep desire for a constructive and collaborative relationship with their principals regarding their professional evaluation" (Zimmerman & Deckert-Pelton, 2003). This constructive and collaborative relationship which the authors found to be lacking can be established by administrators if they provide feedback which is perceived as honest and helpful (Marshall, 2005). To provide thoughtful feedback, principals must devote their full attention to a classroom for an uninterrupted block of time (Zatynski, 2012). If principals are to be the primary evaluators for teachers, then adequate preparation and ample time to establish a valid assessment of the teacher's performance in the classroom are crucial. Unfortunately, studies have found that this training and time allowance is not always adequate (Ingle, Rutledge, & Bishop, 2011; Johnson & Roellke, 1999; Marshall, 2005; Ramirez, Lamphere, Smith, Brown, & Pierceall-Herman, 2011). Hill, Charalambous, & Kraft (2012) described the problem as such:

> We also know from anecdotal evidence that principals may be pressed for time and thus "sample", in a sense, a half hour from a lesson before moving on to their next responsibility. If the reliability of teacher scores is not affected by this sampling, then, at least from a measurement perspective, it is a smart strategy. However, if principals are systematically missing important aspects of instruction because of this time-saving approach, their ratings might not consistently capture the overall quality of instruction occurring in the classroom. (p. 57)

Evaluation by third parties. If training and adequate observational time are obstacles to fair and accurate teacher evaluations, then a logical fix would be to assign the task of evaluations to outside observers who have both adequate training and time to devote to the fidelity of the process. The Measures of Effective teaching (MET) project, undertaken in 2009 and funded by the Bill & Melinda Gates Foundation, was intended to build and test measures of effective teaching to find out how evaluation methods could best be used to help districts and teachers identify effective teaching and improve teacher quality. Over 3,000 teachers across six school districts volunteered to participate in the project, with the goal of determining how to "close the gap between expectations of effective teaching and what is actually happening in classrooms" (http://www.metproject.org). The project examined student survey data, classroom observation data, student achievement data, teacher content knowledge, and teacher perceptions of working conditions.

As one aspect of their investigation, the MET project compared teacher evaluations conducted by principals to those conducted by "peers" – fellow teachers trained in and tasked with observing other teachers. In a sample of 129 administrators using a four-point observation scale, they found that principals rated their own teachers

slightly higher (0.1 points) than principals from other schools and even slightly higher (0.2 points) than peers (Ho & Kane, 2013). These differences may seem quite small, but given the highly compressed data (scores were disproportionately clustered in the middle of the distribution, another common finding in principal evaluations), a 0.1 point difference in mean observation score was equivalent to ten percentile ranks.

In an interesting finding, although administrators scored their own teachers higher in terms of raw score, their rankings were similar to the rankings produced by others outside their schools, with an overall Spearman rank-order correlation of 0.87 (Ho & Kane, 2013). This implies that administrators' scores were not heavily driven by factors outside the lesson, such as a principal's prior impressions of the teacher or personal bias.

Another study (Searles & Kudeki, 1987) concluded similarly that principals (N = 28) and teachers (N = 81) do not have a tendency to rank teachers differently based on observable factors. The authors administered a 100-item questionnaire to both principals and science teachers on the effectiveness of observed science lessons. For 85 of the 100 items, results from Pearson chi-square tests suggested that there was no evidence that principals and teachers classified the observed teachers differently, suggesting principals and teachers tended to rank teachers similarly.

Likewise, Searles & Ng (1982) showed that principals (N = 22) and biology teachers (N = 41) tended to rate biology teachers similarly. On an observational checklist, 79 of 84 items exhibited no statistically significant difference between the two groups.

The reality of teacher evaluation, however, is that although principals may rank teachers similarly to the way their peers might rank them, rankings do not affect a

teacher's employment status to the degree that ratings do. If feedback from principal evaluations is to be useful, it would follow that teacher ratings should be differentiated among the rating categories. Studies have found that this is not always the case. One study addressing teacher evaluations of over 8,000 teachers across five large school districts determined that almost 99 percent of teachers receive ratings of "satisfactory," based on their district's standards when a dichotomous rating system was used (Zatynski, 2012). A separate study showed that principals in 87 percent of schools in a single large urban district of over 600 schools did not issue a single "unsatisfactory" rating between 2003 and 2006 (Toch, 2008).

The MET project found a similar phenomenon in a different district:

> In a study where 127 school administrators and teachers observers each conducted 24 teacher observations, observers rarely used the top or bottom categories ("unsatisfactory" and "advanced") on the four-point observation instrument, which was based on Charlotte Danielson's Framework for Teaching. On any given item, an average of 5 percent of scores were in the bottom category ("unsatisfactory"), while 2 percent of scores were in the top category ("advanced"). The vast majority of scores were in the middle two categories, "basic" and "proficient." (Ho & Kane, 2013)

It is a commonly accepted research practice to collect more data if the accuracy of a phenomenon is in question. With such a compressed rating scale, the accuracy of these teacher ratings would surely be in question. One solution, then, would be to increase the number of observations upon which a teacher's evaluation is based. Another would be to

decompress the rating scale. The MET project came to the first conclusion, by suggesting that a district base a teacher's evaluation on both principal observations and those by outside observers. This allows for the convenience of observations carried out by administrators in the building and with whom the teachers are already familiar, acknowledging that "adding observations by observers from outside a teacher's school to those carried out by a teacher's own administrator can provide an ongoing check against "in-school bias" (Bill and Melinda Gates Foundation, 2013, p. 5).

In any case, effective teacher evaluation relies on the instrument and training processes as much as the observers themselves. One way to align these aspects is to choose an observation instrument that sets clear expectations (Bill and Melinda Gates Foundation, 2012). With an instrument that sets clear expectations, principals can use their contextual knowledge of a teacher in combination with their observations, and outside observers can rely on an outsider perspective.

Reliability and validity of observations. As many teacher evaluation systems, including the one in this study, are at least partially dependent on teacher observations by trained evaluators using a rubric, any discussion of the usefulness of their scores should include a discussion of reliability. It stands to reason that, as with any score, any one teacher observation score is actually comprised of various components, of which error is one.

The MET project found reliability values of 0.14 to 0.37 for single teacher observation scores. In other words, the vast majority of the total variance in scores across all lessons, teachers, and raters -- typically around two-thirds -- is due to factors other

than persistent differences among teachers, such as observer differences, random fluctuation in lesson quality, and other factors (Bill and Melinda Gates Foundation, 2012). In a separate report, MET noted that reliably characterizing a teacher's practice required averaging scores over multiple observations. In their study, the same teacher was often rated differently depending on who performed the observation and which lesson was being observed. The influence of an atypical lesson and unusual observer judgment were reduced with multiple lessons and observers. (Bill and Melinda Gates Foundation, 2012, p. 2). This is not to say that much variance was found to come from observer bias. It was found that less that 10 percent of total score variation actually resulted from some observers consistently scoring higher or lower than others. This underscores the ability to control the effects of over- or under-scoring through effective training.

Although persistent over- or under-scoring didn't seem to be a substantial problem, the project did find that teacher scores vary more among observers than among lessons observed by a single observer. Adding a second observer when multiple observations were to occur increased reliability from 0.66 to 0.72, compared to the alternative of having the same observer conduct multiple observations on the same teacher. In fact, multiple shorter evaluations can be more reliable than longer observations by the same observer, a reliability increase from 0.66 to 0.69 (Bill and Melinda Gates Foundation, 2013, p.5). Other studies have focused heavily on assessing inter-rater reliability for teacher observations, and typical values range from 0.55 to 0.89, with most falling above 0.74 (Gersten, Baker, Haager, & Graves, 2005; Van Tassel-Baska et al., 2006).

One interesting study (Hill et al., 2012) brought up an important point regarding inter-rater reliability – namely, that inter-rater reliability as typically measured is largely influenced by the number of points on the scale. A scale with two score options would typically provide a much higher level of inter-rater agreement than one with ten score options, regardless of the quality of the instrument or the quality of the lesson being observed. The authors referred to the reliance on inter-rater reliability as "misleading," because multiple factors go into an observation's "true" reliability, and inter-rater reliability focuses only on one factor. Hill et al. write that even strong rater agreement does not assure the consistency of teacher scores and may even mask problems with the data. In addition to the number of points on the rating scale, agreement is influenced by the frequency of target behaviors observed and chance. The authors also noted that inter-rater agreement focuses only on the rater as a source of agreement or disagreement and leaves unexamined other sources of variation such as instruments, training methods, and specific scoring designs.

Given these findings, along with the knowledge that school districts do not have unlimited funds for hiring and training observers, some suggestions are provided for maximizing the efficiency of classroom observation. It was found that observations based on the first 15 minutes of lessons were about 60 percent as reliable as full lesson observations, although requiring one-third as much observer time, suggesting that the optimized solution may lie in shorter, more frequent observations with multiple observers (Bill and Melinda Gates Foundation, 2013, p. 19). A logistical issue, however, prevents overreliance on short observations, because some aspects of teaching monitored by

observers don't tend to be observable during short observations – particularly if they are limited to the beginning of a class period (Bill and Melinda Gates Foundation, 2013, p.19).

The most salient point taken away from the research on improving reliability of teacher observations is the importance of requiring observers to be well-trained. Ideally, observers will demonstrate competency, consistency, and accuracy before conducting observations on live lessons.

Determining reliability of a teacher observation instrument is a fairly straightforward process, but determining validity can be more difficult. One way to establish validity of a particular system of evaluation is to analyze the relationship between observation scores and an already-existing measure of teacher effectiveness.

In a study of the Chilean national teacher evaluation system, Santelices and Taut (2011) found that teachers exhibited significant differences in the expected direction between teachers rated as "satisfactory" and "unsatisfactory" on teacher observation measures. Although the data were gathered from a school system outside the United States, the findings are relevant to the current study, as teacher effectiveness variables used were similar to those used in US-based studies.The 58 teachers in the sample were divided into "satisfactory"(n=32) and "unsatisfactory" (n=26) groups, and statistically significant differences were found between the two groups on such disparate measures as student achievement, content knowledge, and teacher observation scores. In particular, they found "especially strong and practically significant differences related to time on

task during lessons, lesson structure, student behavior, and student evaluation materials" (p. 72).

Another study examining criterion-related validity of a teacher observation instrument correlated scores from each instrument subscale with residualized student growth scores in reading performance. The various instrument subscales showed moderate to moderately strong correlations with student growth, ranging from 0.49 to 0.65 (Gersten et al., 2005).

Another approach to establishing validity of a teacher observation system was exhibited in Van Tassel-Baska, et al. (2006). An "expert panel" of professors, scholars, practitioners, and administrators was assembled to rate the instrument's content validity, and intra-class coefficients were calculated to determine expert agreement on the validity of the form. Their agreement on the two dimensions of the validity was .86 for the importance of each item, and .99 for the clarity of language.

Use of and Issues with Value-Added Modeling

Another method of teacher effectiveness measurement included in this study is Value-Added Modeling (VAM). In North Carolina, teachers who teach state-tested subjects receive value-added scores through the Educator Value-Added Assessment System (EVAAS®), designed by SAS Institute Inc. and based on the Tennessee Value-Added Assessment System (TVAAS) (Sanders, Saxton, & Horn, 1997). The primary appeal of VAM to measure educator effectiveness is in its ability to capture student growth, rather than simple student achievement.

Value-added models were brought to education with the intent of measuring the specific contribution of a school (or teacher) to a student's learning. Whereas traditional methods of student achievement measurement focus on proficiency and performance level, VAMs involve a prediction model that controls for, at a minimum, a student's prior achievement. When out-of-context student achievement, such as raw achievement scores or the proportion of students above a particular proficiency bar, are used as the primary quantitative basis for teacher evaluation, teachers can be unfairly blamed or helped simply by having a classroom full of students with a history of low or high achievement, respectively. Value-added models (of which there are many forms) use various degrees of student-, classroom-, and school-level covariates (e.g. student demographics, SES, school size) to attempt to isolate the effect of interest. In general, VAMs rely on the assessment of student growth – students are predicted to perform at a particular level on a standardized test determined by the variables in the prediction model, and any deviation from the prediction is attributed to the teacher (and/or the school).

In this sense, VAMs are distinct from more basic measurements of student growth. Growth models, like VAMs, track the same students over time, measuring their performance at one point in time relative to a previous baseline. However, these simpler growth models implicitly assume that all students learn at a uniform rate and that a teacher (or school) is the only responsible party for growth (Scherrer, 2011). VAMs also track student test score change over time, but in a way that models the added effect of a teacher after controlling for other effects, depending on the VAM used (Timmermans, Doolaards, & de Wolf, 2011).

However, as articulated by Van de Grift, "It is far easier to define the value added of schools than to assess it" (2009, p. 270). Although there are many applications of the value-added concept in educational assessment, one of the more common approaches is a general model from the work of McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004).

The general model. The general VAM described by McCaffrey et al. is restricted in that it only depicts projections for one subject, and contains variables that allow for the inclusion of a single year of test scores. The model, however, can be generalized to any number of consecutive grades or subjects. McCaffrey et al. identify the general model for a student's initial year of testing as:

$$y_{i0} = \mu_0 + \beta_0' x_i + \gamma_{00}' z_{i0} + \sum_{k=1}^{M} \lambda_{i0k}\eta_{0k} + \sum_{j0=1}^{N_0} \phi_{i0j}\theta_j + \epsilon_{i0}$$

In this notation, 0 refers to grade 0, the first year of a student's test history. In general notation, the y_{ig} refers to the test score of student *i* in grade *g*.The μ_g represents the district (or state, or reference unit) mean in grade *g*, and x_i and z_{i0} refer to student-level covariates that are time-invariant (gender, SES, etc.) and time-variant (testing accommodations, setting, etc.), respectively. For the *M* schools in the district, the λ_{i0k} refers to the proportion of the year spent by student i in grade 0 at school *k,* and the η_{0k} refer to the school effects for school *k* in grade 0. Similarly, for the N_0 grade 0 teachers, Φ_{i0j} refers to the proportion of the year that teacher *j* taught student *i* in grade 0, and θ_j

refers to the teacher effect of teacher *j*. The ε_{i0} are normally distributed residual errors for student test scores.

As the authors note, this model can be adapted to fit many types of value-added models by setting any of the above coefficients to zero. In order to project student growth and teacher effect for a future year (year 1 using the above notation), terms can be included that measure the impact of previous years' teachers on a student's projected score. This "alpha" value can be set to 1 to imply that a teacher's effect on student learning continues unabated through a student's career, set to 0 to imply that previous teacher effects have no bearing on future learning, or set to some decaying amount as years progress. In the above model, previous test performance would be included in the student-level time-variant *z* variables.

Types of value-added models. One of the primary concerns in choosing a model, at least on a conceptual level, has been the issue of whether to include student demographic data in the prediction model. Martineau (2006a) states, for example, that fairness toward educators has been an important rationale for implementing VAM and has led the Dallas school district to include large numbers of policy-defined ''fairness'' variables as covariates in the models (p. 252).

Timmermans et al. (2011) categorize commonly-used VAMs into five types, which will be summarized here. The authors point out that because VAMs differ in the types of control variables they employ, they also necessarily differ in the meaning of their output. Any model that doesn't explicitly include a certain variable that affects student

achievement results in an error term that includes the effects that are not accounted for by the variables in the equation. The five types identified are:

1) "Type 0" – This is a model without control variables. This is simply a model that compares the average student performance at a school to the average student performance at the average school in the district. The authors point out that this cannot be considered a true value-added model, as no effects are isolated. This model would be inappropriate to use in practice, as school effect estimates contain many variables outside of the school's control.

2) "Type AA" – The Type AA model is a model which contains controls only for prior student achievement. The VAM used in this project is of this type. The output of this model compares student performance at a school with student performance at the average school among students of similar prior achievement. This model contains the implication that all learning takes place due to the school (or teacher, if applying this model for teacher comparisons).

3) "Type A" – This model is like the Type AA model, but with the inclusion of additional student-level controls. These added controls are typically background characteristics such as gender, ethnicity, or socioeconomic status. This model compares student performance with the performance of other students who are similar in both prior achievement and background. Use of this model would address the concern that students from differing

backgrounds tend to differ not only in beginning achievement level, but also in rate of growth (Linn, 2001; Kupermintz, 2002).

4) "Type B" – These models include both student-level prior achievement and student-level controls and add school-level compositional effects such as average SES and average prior achievement. These approaches compare student performance with the average performance of similar students at schools that are compositionally similar.

5) "Type X" – Type X models are very similar to Type B, but also include nonmalleable school-level controls, such as school size, location, and staffing patterns. These models compare student performance to average similar student performance at schools that are both compositionally and structurally similar.

These five models are conceptually distinct, and Timmermans et al. (2011) performed a series of studies that compared their empirical effects in terms of school/teacher ranking and classification. In a sample of 137 Dutch schools, it was found that all five models showed school effect correlations ranging from .68 to .85, suggesting that they are all measuring a similar underlying construct. In a smaller study (N = 22 middle school mathematics teachers), Hill, Kapitula, & Umland (2011) analyzed teacher effects from models AA, A, B, and X, and found Spearman rank-order correlations between all pairs to be equal to or greater than 0.82. However, results from this study may be questionable due to the small sample size.

The Timmermans et al. (2011) study also analyzed classification effects of the five models by sorting schools into underperforming, average, and overperforming groups based on significant school effect difference from zero. Coefficient Kappa values among the four models (not counting Type 0, that only exhibited slight agreement with the other models) ranged from .34 (Type AA with Type X) to .78 (Type B with Type X).

These findings suggest that the four models tend to classify schools in at least a moderately similar fashion. Progressing from simple to more complex models tended to increase the size of the high- and low-performance groups while shrinking the size of the average groups. More complex models explained more of both the school and student-level variance. For example, going from Model Type AA to A increased explained student-level variance from 17 to 27% and increased explained school-level variance from 35 to 42%. Interestingly, however, going from type B to X added no explained variance.

Another study found conflicting results with regard to the explained variance of Type X versus Type B. In a study of approximately 500 Australian schools (relevant to US schools due to its examination of similar value-added models), Keeves, Hungi, and Afrassa (2005) found that the Type X model, by including non-malleable school-level effects, increased variance explained in student scores by 13 percentage points over a Type B model. To this point, Timmermans et al. (2011) noted that conceptually, model Type B is preferable to Type X despite the additional explained variance if the purpose of modeling is to identify weak schools, regardless of their size. Including nonmalleable

factors such as school size in the model only identify as underperforming those schools which underperform compared to schools *of similar size* (and similar on other included nonmalleable factors). If a school is to be labeled as weak, the authors argue, it should not only be within the context of structurally similar schools.

The Keeves et al. (2005) study found similar results to the Timmermans et al. study when comparing unexplained student-level variance from a Type A model (53.1%) to the same from a Type B model (46.4%). These studies agree that the addition of school-level compositional covariates produces more useful results.

Criticisms of VAM. Even though the general aim of VAM is a conceptual improvement over the use of student-based outcome measures without controls as teacher evaluation tools, there remains much criticism about the use and overreliance on VAM. These criticisms exist on both an empirical and a conceptual level, both of which will be addressed in this section.

Van de Grift (2009) enumerates four empirical issues with the use of VAM for teacher evaluation: First, there is typically too much missing data to claim valid teacher effect estimates. In his study, the author found that 34% of students tested in a school district in the Netherlands were no longer in the sample six years later. In addition, 7% of students in Year 8 of his sample were not in the sample two years previous, and 11% were not in the sample four years previous. This would not be a significant problem if the students missing from the sample were random, but his second issue addresses this concern.

The second issue raised is that the missing data are not random. The author found in his study that 32% of students who disappear from a cohort do so due to retention. This alone has the effect of removing struggling students from the sample for future teachers, but there also seems to be a demographic effect. In an analysis of the retained students, it was shown that 27% of students from Turkish or Moroccan descent were retained at some point, compared to 17% of low-SES students and 11% of non-Turkish or Moroccan students.

Third, the results are unstable. In his study of 33 Dutch elementary schools over six years, no schools persisted in the over- or under-achievement category (defined by their school effect being significantly different from zero) over all six years. 29% of schools persisted in the middle category for all six years. The other 71% of schools all spent at least one year in the over- or under-achievement category. Again, though the study was performed on Dutch schools, the concepts of student mobility, nonrandom assignment of students to schools, and retention are applicable to US-based schools as well.

Much other work has been done on the stability of value added estimates. For example, Koedel and Betts (2007) used VAM to identify effective teachers, and found that only 35% of teachers in the top 20% in one academic year remained in the top 20% the next year. In a large study of many uses of VAM, McCaffrey, Sass, Lockwood, & Mihaly (2009) found that year-to-year correlations of teacher effect tend to be in the 0.2 – 0.5 range for elementary teachers and 0.3 to 0.7 for middle school teachers. In an

interesting note, McCaffrey et al. (2009) point out that correlations of 0.3 -- 0.7 in year-to-year performance are similar to the year-to-year performance measures of other professionals, such as insurance salesmen and baseball players, whose output is measured directly. In addition, in a large study of the effect of measurement error on VAM, the Bill and Melinda Gates Foundation clarified an alternate interpretation of the relatively low year-to-year teacher effect correlations: "if student achievement gains are an imperfect measure of a teacher's underlying value-added in one year, they also will be subject to error in any other year. The year-to-year correlation is diminished by the fact that there is measurement error in both years" (Bill and Melinda Gates Foundation, 2012, p. 44).

Finally, teacher rankings vary depending on the VAM chosen. This criticism is addressed above, in Timmermans et al. (2011), who found that when comparing classification tendencies across five VAMs, the level of agreement as measured by Kappa can be as low as 0.34.

Other studies have criticized VAM for drawing conclusions about teacher or school contributions to student learning based on effect scores that are measuring more than the intended effect. One particularly noteworthy example comes from Rothstein (2010) who, after developing falsification tests for three commonly-used VAMs, was able to show fifth-grade teacher effects as a significant predictor of fourth-grade student score gains, a finding that should be impossible in reality. The author discusses this as an indication that VAM teacher effects contain non-modeled information about the students,

such as the non-random sorting of students to schools and teachers, and that these effects should be interpreted with caution.

McCaffrey et al. (2004) conducted a simulation study where various VAMs were employed to determine teacher effects in classrooms with different numbers of faster- and slower-gain students. The authors generated student data, including mean gain rates for each of two groups of students. The data was generated to compare the VAM scores of teachers who taught in a theoretical classroom entirely populated by fast-gaining students to teachers who taught in a theoretical classroom entirely populated with slow-gaining students, and theoretical classrooms with various ratios of the two. The difference in mean gain rates between groups was one gain score standard deviation unit. They found that across all models, teacher effects were moderately correlated (approximately 0.45) with the classroom percentage of fast gainers, suggesting that the estimated teacher effects were actually measuring the student gain speed independent of the teacher, to some degree.

Although VAMs used in teacher evaluation depend on standardized test scores, it makes sense that the test itself, and its associated measurement error, would have an impact on estimated teacher effects. In one study, Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007) found that math teachers' value-added scores, across four different types of VAMs, were extremely sensitive to the particular math assessment used. In their sample of 34 teachers receiving effect scores from two different math assessments across three years, the lowest correlation between model types (holding

assessment constant) was 0.49. The highest correlation between differing-assessment teacher effects was 0.46. These findings show that the particular model chosen to estimate effects makes less of a difference in teacher effect score than the math assessment used.

Papay (2012) performed a similar study and obtained similar results. In a sample of approximately 700 teachers across nine VAMs, the author found Spearman correlations ranged from 0.15 to 0.58 of teacher effects across three different, but conceptually similar reading assessments. He found that measurement error in the tests and the timing of the tests are the primary sources of variation in teacher effects. He notes that the correlations were all statistically significant and most are moderately sized, suggesting that, on average, teachers whose students perform well on one test tend to perform well on other tests. However, they were sufficiently low that they produced substantially different classifications of many individual teachers (p. 179).

As mentioned above, some researchers have conceptual arguments with the use of value-added assessment of teachers in addition to the empirical ones. Hill, Kapitula, and Umland (2011) stressed the importance of considering value-added scores from more than a statistical standpoint. They pointed out that statistical tests of VAM scores provide benefit but do little for advancing the larger issue of VAM validity in terms of how they are used in assessment systems.

Harris (2009) addressed his concerns by enumerating the assumptions inherent in VAM: (1) School administration and teamwork among teachers do not have a significant

impact on student achievement; (2) Controlling for previous achievement levels is sufficient to account for the impact of past school resources; (3) Students' contributions to their own achievement can be measured with student fixed effects that account for the nonrandom assignment of students to teachers; (4) A one-point increase in test scores represents the same amount of learning regardless of the students' initial level of achievement or the test year; and (5) Teachers are equally effective with all types of students. He claimed that none of the five assumptions are supported by research in schools, and that violation of these assumptions leads to serious issues when results of VAMs are used to reward or punish teachers.

Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein (2012) offered a similar criticism by pointing out that, in order for a VAM to effectively measure teacher quality, one must assume that, "student learning is measured well by a given test, is influenced by the teacher alone, and is independent from the growth of classmates and other aspects of the classroom context," and that "none of these assumptions is well supported by current evidence" (p. 8).

Scherrer (2011) raised the issue of validity with a reliance on VAM teacher effects by stating that the standardized tests used in calculating teacher effect estimates don't typically capture the true stated goals of student learning, such as critical thinking skills. He likened teachers aiming to teach to standardized tests while ignoring critical thinking skills to picking low-hanging fruit.

Scherrer also raised the issue of the "rational teacher" – a teacher who acts in their own self-interest by not doing anything to create learning for students whose scores will not contribute to their own VAM, and who would refuse to assist other teachers in educating students who will count for the other teacher's VAM. This would happen because VAM is normative – teachers are compared to one another and not to criteria. Therefore, one teacher's gain is necessarily another teacher's loss. The author illustrated an example where two teachers on a team have complementary skills and decide to team teach. Teacher A is strong in math instruction and weaker in reading instruction, whereas teacher B is strong in reading instruction and weaker in math instruction. To combine their strengths, Teacher A instructs both classes in math, while Teacher B instructs both classes in reading. Even though this would seem to be a beneficial arrangement for the students, Papay (2012)argues that a rational teacher under a potential reward or punishment system under VAM would then have an incentive to provide weaker instruction to the student from the other teacher's class. In a note relevant to this project, the VAM used in North Carolina attempts to correct for this by performing verifications to ensure that student scores are attributed to the teacher who actually taught the tested subject, not necessarily the teacher whose class a student is in.

Another common validity issue raised by researchers (e.g. Marder, 2012; Martineau, 2006a; Scherrer, 2011) is that the skills taught in consecutive grades are often orthogonal (i.e. statistically independent) to one another. If the knowledge and skills in, say, 4^{th} grade are completely orthogonal to those required in 5^{th} grade, then there are problems in suggesting that the teacher is completely responsible for the knowledge

tested. Students may come in with more or less knowledge about the skills to be tested in the new grade, and there is no way to capture it. Consider a student being tested on arithmetic in year one, and being tested on geometry in year two. If the student has no background in geometry before year two, and arithmetic and geometry are orthogonal (an assumption made purely for this example), then a class of students may have a completely different ability distribution on arithmetic and geometry, and the model wouldn't be able to accurately predict a student's geometry performance based on arithmetic performance. In essence, there is no true established baseline for students in that case.

This reliance on testing disconnected skill sets leads to teachers having a problem "buying in" to VAM scores. Buy-in also suffers based on the fact that the calculations that lead to the scores are effectively a "black box" (Gitomer, 2011). Similar to the frustration of teachers not understanding how their scores are calculated, Collins and Amrein-Beardsley (2013) found that, in a nationwide survey of VAM use among state education agencies, no state representatives had plans to use VAM data in a formative way. They point out that, to date, no research has shown that providing teachers access to value-added data produces positive educational outcomes. Although this may not be a concern if the only use of an evaluation instrument is identifying high- and low-performing teachers, this is a major concern if evaluation data is to be used constructively.

VAM used in this project. School districts in the state of North Carolina use, as one method of teacher evaluation, the Education Value-Added Assessment System (EVAAS®), developed by SAS Institute Inc. (Wright, White, Sanders, & Rivers, 2010). This particular VAM would be defined as a Type AA model, as it only controls for a student's prior achievement at the student level and contains no controls at the school level. EVAAS® is based on the TVAAS model (Sanders, Saxton, & Horn, 1997). Despite only controlling for prior achievement, a case can be made that EVAAS® is the most robust and efficient model currently in use due to the model's method of handling missing data and the software's ability to handle large-scale analyses (Amrein-Beardsley, 2008).

Student scores on vertically-scaled assessments are the most straightforward to use in VAM, because measuring student growth is as simple as calculating a gain score by subtracting a student's current score from her previous score. EVAAS®, however, is also robust to non-vertically scaled tests by using different models for the two different types of test. For non-vertically scaled assessments, the Univariate Response Model (URM) is used, which uses predicted scale scores rather than mean gains. In the Multivariate response model (MRM), scale scores are converted to normal curve equivalent (NCE) scores. Mean gains are computed using the NCE scores.

As previously mentioned, EVAAS® does not require vertically scored assessments in order to produce teacher effect estimates. Rather, EVAAS® has three criteria for the scales used in EVAAS® analyses: (1) they must be highly correlated with

state standards; (2) they must assess performance across the full range of scores; and, (3) they must be reliable (Wright et al., 2010).

Amrein-Beardsley (2008) pointed out that a commonly-touted feature of EVAAS® is the way the software handles the typically large amounts of missing data encountered in longitudinal growth analysis. Rather than imputing the missing data, EVAAS® accounts for missing data by using the correlation between current and previous student scores in the non-missing data to estimate means for the current and previous scores. Then, the difference between the two means is an estimate of the average gain for that set of students.

As mentioned above, EVAAS® uses two distinct models to handle two sets of data conditions: the Multivariate Response Model (MRM, a linear mixed model) is used when the test data are scaled to expect reasonable continuation of progress, and when the data have been collected over many schools and/or districts; in North Carolina, the MRM is used with elementary and middle school math and reading End-of-Grade exams. The Univariate Response Model (URM), that is similar to an ANCOVA, is used when the criteria for the MRM are not met. In North Carolina, the URM is calculated for secondary End-of-Course exams, 5^{th}-grade and 8^{th} grade science exams, Career and Technical Education exams, and common exams. The MRM is the preferred model (Wright et al., 2010) because the growth standard is anchored to a comparison population rather than normed each year, theoretically allowing all schools/teachers to meet or exceed expected

growth. In the MRM, the value-added mean gain is determined by comparing an estimate of current mean achievement level to an estimate of a past mean achievement level.

Teacher effect estimates for 5th grade science, 8th grade science, and middle school End-of-Course tests such as Algebra I are calculated using the univariate response model (URM). The URM is used when tests are not given in consecutive years and thus not vertically scaled, and when data from multiple tests is used (Wright et al., 2010). The URM, unlike the MRM, sets a new growth standard annually, based on that year's student performance, and teacher effects and indices are based on each year's norms. With the URM, it is not theoretically possible for all teachers to score at or above expected growth in a given year. It should be noted that both the URM and MRM are proprietary models and formulations are therefore not available.

In addition, EVAAS® relies on the intra-student correlation matrix to increase the accuracy of prediction. When student scores from multiple previous subject and grade assessments are taken into account, the amount of variance in student scores explained is increased and predications can be calculated for subject tests for which there exist no previous analogous data (Ballou, Sanders, & Wright, 2004). For this reason, EVAAS® requires at least three previous test scores for a student before a prediction will be calculated using the EVAAS® URM (Sanders, 2006; Wright et al., 2010).

Although EVAAS® has been described as the most recognized and widely-implemented VAM (Amrein-Beardsley, 2008), researchers have noted other shortcomings beyond the general criticisms of VAM. In one study, McCaffrey et al.

(2004) compared their general VAM model to the layered model (a version of the TVAAS model) and found evidence (likelihood ratio statistic $p < .001$) that, unlike as modeled in TVAAS, teacher effects don't persist unabated from year to year, but rather exponentially decay.

The most common criticism of EVAAS® specifically is that the model does not control for student- or school-level demographic or contextual factors. Linn (2001) suggests that controlling for student prior achievement goes a long way in leveling the playing field for teachers of students from various backgrounds, but it doesn't go far enough. He claims that adjustments for differences in student achievement "do not preclude the possibility that students from different SES backgrounds will have different levels of support for learning and differential access to enrichment experiences outside of school" (p.17), and that these differences could cause systematic bias in the estimation of teacher effects.

In a similar criticism, Kupermintz (2003) states that TVAAS errs by using student prior achievement as a proxy for all other background variables, and cites several studies (Ballou, 2002; Ladd & Walsh, 2002; Sanders, 2006) that address the correlation of value-added scores and school SES, the non-random assignment of students to teachers, and the teacher classification issues that affect teacher effect estimates when SES is not explicitly modeled. All of these arguments stem from the concept that student background factors affect not only a student's starting point, but also their rate of learning.

Ballou, Sanders, and Wright (2004) directly addressed these criticisms both conceptually and empirically, first pointing out that if assignment of students to teachers and teachers to schools is not random, then demographic and SES variables become proxies for school and teacher quality. Because these variables are correlated with unmeasured fluctuations in school quality, these coefficients would contain part of what should be measured by residuals, and predictors of effectiveness would be biased toward zero. Therefore, the authors state, that when low-SES students don't gain as quickly, it's difficult to know whether that's the effect of the background or the effect of the school.

In order to test the criticisms regarding the lack of demographic controls in TVAAS, the authors modified TVAAS in three ways: (1) by adding student-level covariates, (2) by adding student-level covariates and school-level FRL, and (3) by adding student-level covariates and grade-within-school-level FRL, for a total of four models. The authors calculated all six pairwise correlations among the four models for each of 15 grade/subject combinations, and reported the smallest pairwise correlation for each grade/subject combination. Correlations between all pairs of models across three subjects and six grade levels ranged from 0.53 (7^{th} grade Language Arts) to 0.98 (4^{th} grade Language Arts), with all but one being above 0.75. These findings suggest that teacher effects tend to be similar regardless of whether one controls for demographic differences.

When testing the effect of model differences on teachers in high-poverty situations, Ballou et al. (2004) found an average difference in math teacher effect of 0.26

for a teacher of a 100% FRL classroom when student-level demographic controls were included, with the cut score for being classified as "exceptional" at 1.5. On the other hand, adding a school-level FRL variable increased that difference to 2.5 for the same group – about ten times greater than using student-level FRL alone. Based on these results, the implications of including school effects in a teacher-level VAM are vast, while including student-level demographic data may not be substantial.

Ballou et al. (2004) also addressed classification issues and their sensitivity to demographic inclusion. In terms of student-level variables, little effect was found on teacher effect scores. When classifying teachers into "below proficient," "proficient," and "exceptional," agreement between the four TVAAS-based models was 2.7 times more likely than disagreement in reading, 3.5 times more likely in ELA, and 8.5 times more likely in math. Overall, for the addition of school- and grade-level FRL variable inclusion, substantial differences in classification were found, but the estimates of the coefficients were unstable and had large standard errors. They swung from positive to negative across grade levels and the authors questioned the accuracy of the process.

Ballou et al. (2004) offer potential counters to those surprised that including student-level covariates doesn't greatly affect estimates, and disproved all but one:

1) Teachers' classrooms don't vary as much as assumed – This was disproved by showing the mean proportion of students receiving free or reduced lunch in their sample was 47% with an SD of 23%.

2) The impact of SES and demographics doesn't matter as much as expected – this is opposed by the large effect of school- and grade-level FRL.

3) SES and demographic variables don't provide much unique information beyond what is already contained in the covariance matrix of test scores. This is true, because intra-student scores correlate at 0.6 to 0.7 across subjects within a grade, and at 0.8 across grades within a subject. When this information is modeled (as it is in TVAAS), student-level demographics don't add much to the estimation. When intra-student correlations are not used, demographic variables change the estimates substantially. In this sense, the key of TVAAS lies in using a student's entire testing history (up to 5 years' worth) for estimation, rather than only tests in the same subject.

McCaffrey et al. (2004) agreed with the third point above, as they found that in a scenario in which there were homogeneous distributions of fast and slow gainers across classrooms, models that didn't model intra-student correlation showed high correlation (.79) between teacher effects and fast-gainer percentage. This correlation dropped to 0.47 when intra-student correlations were modeled.

In a conceptual show of support for the explicit exclusion of demographic variables, Martineau (2006a) wrote that, although fairness is a real concern, if such variables are included in a VAM, educators who are charged with teaching students whose demographic variables would suggest a slower rate of growth would then only be held accountable for maintaining this slow rate. This essentially lowers the bar for

teachers of students whose backgrounds would suggest a slower rate of growth – an ethical concern.

Combining Measures of Teacher Effectiveness

Rothstein (2012) pointed out that, whereas much research has been done on the development and validation of teacher effectiveness measures, "relatively little attention has been paid to the design of policies that will use the new measures to improve educational outcomes" (p. 2). Even if there is disagreement on the best methods for assessing teacher quality and the degree to which teacher quality directly impacts student outcomes, evidence suggests that teachers do matter, and that there is variance in teacher quality (Papay, 2012). Sanders (2006) points out that two consecutive years of ineffective teachers can leave a detectable negative impact on future growth beyond two years of "catch-up" with effective teachers. Hill et al. (2011) noted that teacher impact routinely explains a higher percentage of variance in student achievement than do school- and system-level factors, and other studies have shown that teacher quality strongly impacts such variables as student engagement, student focus, and student performance on other, higher-level assessments (Gersten et al, 2005).

When discussing what qualities comprise effective teaching, then, it is necessary to consider two findings of Santelices and Taut (2011). The authors compared 58 teachers who had been rated either 'unsatisfactory' or 'outstanding' by the national teacher evaluation system of Chile by analyzing four data points for each teacher: student test score growth, classroom observations, a teacher content knowledge assessment, and expert ratings of a teaching materials binder. First, when comparing the student

performance of ‘unsatisfactory’ and ‘outstanding’ teachers, the authors did not find statistically significant differences in student pre-test scores, but did observe statistically significant differences in the student post-test scores of these classrooms (p. 84). Second, results showed that important differences between ‘outstanding’ and ‘unsatisfactory’ teachers were concentrated in their observable teaching practices related to lesson structure, student behavior, design of classroom assessment materials, and their ability to keep students on task most of the time (p. 88). The authors go on to say that the source of primary variance between “outstanding” and “unsatisfactory” teachers was the difference in overall observation scores.

A similar, much earlier profile of an outstanding teacher was constructed in a study by Searles and Kudeki (1987). The study described the successful teacher as one who is able to maintain a respectful classroom atmosphere, who is a subject matter expert, who presents materials in a clear manner, who is concerned with the progress of students, who exhibits creativity and resourcefulness, and one who differentiates instruction according to the needs of the students.

Support for a composite measure. As noted in Stronge et al. (2011), teacher quality is a complex phenomenon, and there is little consensus on what it is or how to measure it. In fact, there is considerable debate as to whether one should judge teacher effectiveness based on teacher inputs, the teaching process, the product of teaching, or a composite of these elements.

Hill et al. (2011) argued strongly in favor of a composite. They emphasized the importance of going beyond simply writing instruments for teacher evaluation and rather

creating more in-depth observational systems. These systems would include hiring criteria for raters, rigorous training protocols, robust scoring designs, and quality instruments.

The MET project echoes this view, suggesting that many commonly-accepted measures of teacher effectiveness contribute to fair and accurate assessment of teaching quality. They found that teacher observation scores were strongly related to student achievement in Math and English, and that the relationships between observation scores and student achievement gains were stronger when combined with other information, such as previous student achievement and VAM. Finally, they found that combining measures led to improved reliability and predictive power for future achievement (Bill and Melinda Gates Foundation, 2012, p. 10). Other research provides strong evidence for the use of multiple measures in a teacher evaluation system as well. Jacob and Lefgrin (2008) found that both teacher observations and VAM were statistically significant predictors of future performance.

The use of a combined measure has exhibited content validity as well. Students of teachers scoring high on a combined measure showed larger performance gains on tests of conceptual understanding, higher levels of effort, and greater class enjoyment than students of teachers scoring lower (Bill and Melinda Gates Foundation, 2012, p. 13).

When using multiple measures in a system of evaluation, many factors must be addressed. Darling-Hammond et al. (2012) argued that, in order for a system to be successful, a district must use multiple observations across the year, conducted by expert evaluators looking at multiple sources of data. Staiger & Rockoff (2010) pointed out that

high-quality information about teacher effectiveness, taken from multiple sources, can be effective in increasing outcomes through hiring and firing practices, performance-based pay, and targeted professional development. In their "State of the States 2012" (2013) report, the National Teacher Quality Foundation suggested that states use multiple measures of student learning to measure teacher effectiveness, specifically highlighting evidence of student learning observed during classroom observations. Jacob & Lefgrin (2008) pointed out that both principal evaluations of teacher effectiveness and VAM are better indicators of teacher quality than traditional methods of teacher tenure and retention assessment, that is, experience and education level.

The call for elimination of single-source evaluation systems in favor of a multifaceted approach to teacher evaluation is echoed throughout the literature. Teddlie, Creemers, Kyriakides, Muijs, and Yu (2006) stated that, "if teacher and school improvement are important goals to schools, or to researchers working in those schools, then different data systems (beyond state-controlled teacher evaluation) are required" (p. 567). Similarly, Stronge et al. (2011) suggested that, "no single data source is valid or feasible for all teachers in all school districts" (p. 252).

With composite measures of teacher effectiveness, teachers who are better at helping students learn can be identified and relatively accurate predictions can be made about teacher performance. If composite measures of teacher effectiveness are attractive, then what aspects of teacher quality should go in to the score? As Teddlie et al. (2006) stated, "having one monolithic dimension, which could only be labeled ''teacher

effectiveness,'' certainly does not adequately capture the essence of the construct" (p. 563).

Elements of a composite measure. Johnson (1997) suggested three constructs be considered: the teacher as person, the teaching process, and the teaching product. Johnson also described six key indicators that were common among teachers and could be interpreted as subscales of the three categories: teacher as subject matter expert, teacher as caring, teacher as exhibiting classroom control, teacher as interactive in communication, students as on-task and attentive or engaged, and student progress and achievement.

The MET project suggested combining observation scores and student achievement gains with student feedback. This combination of assessments was found to lead to increased stability in teacher effectiveness scores (Bill and Melinda Gates Foundation, 2013, p. 5). One of the MET Project reports stated that teachers shouldn't be asked to expend effort to improve something that doesn't help them achieve better outcomes for their students. If a measure is to be included in formal evaluation, they argued, then it should be shown that teachers who perform better on that measure are generally more effective in improving student outcomes (p. 15).

Despite the amount of research appealing for a combined measure including both teacher variables and student outcomes to determine teacher effectiveness, North Carolina (among other states) uses a combination of teacher inputs only to determine pay: teacher experience, advanced degree status (which has since been removed from the 2014

budget), and National Board Certification status. Research has been mixed regarding the impact of these measures on teaching effectiveness.

In terms of teacher experience, Rice (2010), for example, conducted a longitudinal study of over 40 years of teacher effectiveness data and concluded that experience appears to have an positive impact on teacher effectiveness in the first five years of a teacher's career (particularly in elementary and middle schools), and levels off after that, eventually leading to a slight decline after 25 years. The same result was found for a large urban school district using EVAAS as the teacher effectiveness measure for 1507 teachers. Effect scores rose sharply for the first five years and leveled off, with a dip around 25 years (Penny & Ward, 2012).

According to studies on the impact of Master's degree attainment on teacher effectiveness (Copur-Gencturk, Hug, & Lubienski, 2014; Lease & Garrison, 2008), advanced degrees appear to have a positive impact on teacher knowledge and classroom practices, but no impact on outside evaluation scores, student reports of effectiveness, or student outcomes. This result was echoed by Penny and Ward (2012) who found that advanced degree status was not a significant predictor of EVAAS teacher effect. Similarly, research on the impact of National Board Certification status (Goldhaber, Perry, & Anthony, 2004; Rouse & Hollomon, 2005; Stronge, Ward, Tucker, Hindman, McColsky, & Howard, 2007) suggested that more effective teachers tend to apply for NBPTS Certification more often than less effective teachers, but that the certification itself does not have a substantial impact on classroom outcomes. As with advanced degree status, Stronge et al. (2007) found that NBPTS certified teachers tend to score

more highly on measures of dispositional and instructional variables, but not on measures of student learning. Penny and Ward (2012) found that National Board Certification was a significant predictor of EVAAS teacher effect in a multiple regression model.

Effective teaching can be measured, and there are many processes through which to do it. Although much research has been done to assess the mechanics, reliability, and effectiveness of such measures, little research has been done on the most effective way to weight them when establishing a composite score. The aim of this project was to contribute to that discussion by examining and comparing the relationships among three methods of teacher effectiveness measurement.

CHAPTER III

METHODS

Background

Winston-Salem/Forsyth County Schools (WS/FCS) is a large urban school district in North Carolina. Within this district, 16 schools (12 elementary and 4 middle) have been designated as STAR3 schools – recipients of a federal Teacher Incentive Fund (TIF) grant. TIF grants are awarded to districts to fund projects with the goal of exploring alternative, merit-based compensation systems for teachers. To be eligible for the grant, districts must base merit pay on teacher evaluation systems that "differentiate effectiveness using multiple rating categories that take into account student achievement growth as a significant factor, as well as classroom observations conducted at least twice during the school year" (Teacher Incentive Fund, 2010, p. 1).

Like all other North Carolina public school teachers, teachers in STAR3 schools are subject to annual evaluations conducted by their principals. These state-mandated evaluations consist of, among other factors, one to four observations per year using the Standards on the Rubric for Evaluating North Carolina Teachers (McRel, 2009). In addition, many teachers in grades 4-8 in WS/FCS receive individual value-added scores through North Carolina's Education Value-Added Assessment System (EVAAS®). Further, teachers in STAR3 schools are subject to two additional observations per year by

trained, full-time observers using a rubric based on the "Teach" section of Washington, DC Public Schools' IMPACT model (see Appendix A). This study is intended to set the stage for further investigation of the use of various teacher evaluation measures (TEMs) in North Carolina and the effects of certain school- and teacher-level variables on classification and score. As the data collected are from a highly specific subset of schools and teachers, the project is framed as a case study and purely descriptive of the particular sample.

Description of Sample

School characteristics. This study examined the relationships among effectiveness measures from teachers at the 16 STAR3 schools in WS/FCS in the 2011-12 school year. In Table 1, the 16 schools and their demographic characteristics are displayed. As can be seen in the tables, all 16 schools have very high (>80%) FRL percentage, but the schools have a wide range (15.69% - 52.35%) of ELL students. Additionally, all 16 schools have performance composites below the district average, although their campus growth composites were higher.

Table 1

Demographic Characteristics of STAR3 Project Schools Compared to District

School	# of Students	Free and Reduced Lunch %	ELL %
Ashley Elem	486	90.25	23.03
Diggs-Latham Elem	410	96.01	41.90
Easton Elem	557	97.89	49.21
Gibson Elem	731	88.31	23.93
Griffith Elem	551	89.39	35.07
Hall-Woodward Elem	762	96.61	52.35
Hill Middle	281	94.72	33.10
Kimberley Park Elem	274	98.18	15.69
Konnoak Elem	684	92.31	31.64
Middle Fork Elem	365	88.83	28.34
Mineral Springs Middle	491	94.50	19.14
North Hills Elem	369	99.19	24.86
Old Town Elem	591	95.85	47.01
Philo Middle	258	96.48	28.91
South Fork Elem	568	86.59	34.49
Wiley Middle	462	80.97	25.79
WS/FCS District	52,606	54.58	12.30

Table 2

Student Achievement Measures for STAR3 Project Schools Compared to District

School	2012 Performance Composite	2012 EVAAS School Composite Index
Ashley Elem	41.4	-3.1
Diggs-Latham Elem	57.7	0.0
Easton Elem	50.0	-2.1
Gibson Elem	61.9	-0.5
Griffith Elem	67.1	-3.2
Hall-Woodward Elem	62.7	-3.3
Hill Middle	45.6	-4.1
Kimberley Park Elem	55.0	-1.5
Konnoak Elem	54.8	-6.8
Middle Fork Elem	68.1	0.4
Mineral Springs Middle	56.7	-6.7
North Hills Elem	55.0	-1.5
Old Town Elem	66.2	-6.0
Philo Middle	53.1	-0.5
South Fork Elem	58.3	-0.2
Wiley Middle	57.6	-5.3
WS/FCS District	74.4	-8.0

Teacher Characteristics. There were 137 teachers for whom scores on all three assessments existed, and these teachers comprised the sample in this study. Although states take different approaches to aligning teacher effectiveness measurement and teacher compensation, North Carolina (like many other states), uses a simple formula when determining teacher pay. Teacher salaries, with some degree of district-by-district supplemental variation, are based on a teacher's years of experience, their attainment of

degrees beyond the bachelor's, and whether or not they are certified by the National Board for Professional Teaching Standards (NBPTS).

Table 3 shows the descriptive statistics of these pay variables for this sample (N = 137 teachers). The vast majority of these teachers were not Nationally Board certified, and approximately 1/3 of them held an advanced degree. There was a wide range of experience, with the average teacher having taught for 11.37 years.

Table 3

Selected Characteristics of Sample Teachers Compared to District

Characteristic	STAR3 Schools	WS/FCS District
NBPTS Certified	5%	11%
Hold advanced degree	30%	34%
Minimum Experience	0 years	0 years
Maximum Experience	38 years	Not Avail.
Mean Experience	11.37 years	14.8 years

Instruments

North Carolina Teacher Evaluation System. The NC TES assessment consists of 25 items across five standards (North Carolina Teacher Evaluation Process, 2012, see Appendix B). Some of these 25 items are scored based on classroom observations, some are scored based on the collection of artifacts, and some are scored based on a combination of the two. Principals (and in some cases assistant principals) are responsible for the observations, artifact collection, and ultimate evaluation of teachers on the instrument. Each item is scored on a five-point ordinal scale, with labels: (1) Not

Demonstrated, (2) Developing, (3) Proficient, (4) Accomplished, and (5) Distinguished. The standards are as follows:

Standard I: Teachers demonstrate leadership (5 items)

Standard II: Teachers establish a respectful environment for a diverse population of students (5 items)

Standard III: Teachers know the content they teach (4 items)

Standard IV: Teachers facilitate learning for their students (8 items)

Standard V: Teachers reflect on their practice (3 items)

In North Carolina, the variability of teacher scores on this instrument for 2011-12 was not large. The vast majority of teachers were rated either "Proficient" or "Accomplished" on all standards (see Table 4). As with the STAR3 observation tool, there is no published information on the reliability or validity of the NC TES Standards. Each teacher is only observed by one individual, so inter-rater reliability is not available, and raw data on items within standards are not available.

Table 4

Proportion of Teachers Receiving Each Rating by Standard on the NC TES, 2011-12

Standard	Not Demonstrated	Developing	Proficient	Accomplished	Distinguished
I	0.1	1.7	36.3	48.4	13.6
II	0.2	2.4	38.7	48.3	10.4
III	0.2	2.6	47.1	41.4	8.7
IV	0.0	2.1	37.6	51.4	8.7
V	0.2	2.6	48.1	39.8	9.3

Source: North Carolina Department of Public Instruction

Depending on the teacher's years of experience, these scores were based on one to four classroom observations per school year and the collection of artifacts. Career status teachers (those with tenure) receive two informal 20-minute observations and one formal 45-minute observation by administrators. Probationary status teachers (those without tenure) receive three formal 45-minute observations and one informal 20-minute observation. In STAR3 schools, evaluation score data existed for all teachers on standards I and IV, and for 84 teachers on all five standards. This study focused on standards I and IV, as all teachers were assessed on these standards. Teachers who were assessed on standards II, III, and V were primarily new teachers. Including these standards would have allowed only new teachers to be analyzed, which was not the intent of the study.

EVAAS®. In all NC public schools, teachers who teach subjects for which there is an eligible end-of-year examination (including End-of-Grade, End-of-Course, CTE,

and Common Exams), and for whom at least 10 students take the test, receive a value-added score through the Education Value-Added Assessment System (EVAAS®), designed by SAS Institute, Inc.

Most 4th grade teachers in North Carolina public schools receive three EVAAS scores: a reading score, a math score, and a composite score. 5th grade teachers receive four EVAAS scores: reading, math, science, and composite. Middle school math and language arts teachers receive a subject score and a composite score, as do 8th-grade science teachers. All elementary and middle school reading and math EVAAS teacher effects are calculated with the MRM model while 5th and 8th grade science teacher effects are based on the URM model.

All teachers who receive EVAAS® scores are evaluated based on their composite scores, regardless of the number of scores comprising the composite. This study focused on the relationship of the teacher evaluation instrument and observation scores with the EVAAS® composite, as all teachers receive composite scores, and since composite scores are more conceptually comparable with the STAR3 observations and NC TES ratings, which are not subject-specific.

In STAR3 schools in 2011-12, 137 elementary and middle school teachers received EVAAS® composite scores, and these teachers comprise the sample in this study. Table 5 displays the tests that comprise the calculation of teacher effect indices used in the current sample.

Table 5

EVAAS® Teacher Effect Composite Components by Grade

		Composite Components	
Grade	No. of Teachers	MRM	URM
4	49	Math, ELA	None
5	46	Math, ELA	Science
6	15	Math, ELA	None
7	13	Math, ELA	None
8	14	Math, ELA	Algebra I, Science
Total	137	--	--

STAR3 observations. Teachers at STAR3 schools, because of the nature of the grant project, are evaluated by an extra measure of teacher effectiveness. The primary system for teacher evaluation (NC TES) used in the district did not contain an observation protocol that was as detailed as administrators felt was appropriate for the project's goals, so a new observation tool was selected.

After conducting research into observation tools used in other large urban school districts, the "Teach" portion of the Washington, DC Public School System's IMPACT evaluation model was chosen. From a face validity standpoint, the instrument seemed to be broadly applicable, in line with the district's views on effective teaching, and clear in its descriptions. To date, there has been no reliability information published about the tool, nor any published research on the validity of the instrument. In 2011, a conference call between WS/FCS officials and DCPS representatives revealed their inter-rater reliability of 0.80 with full-time observers. In addition, DCPS observation scores showed a 0.34 correlation with value-added scores for teachers (District of Columbia Public

Schools, personal communication, 2011). As described below, inter-rater reliability was calculated for the sample used in this research.

All "core" teachers – defined in STAR3 as teachers for whom value-added scores are generated – receive two third-party observations per year by trained observers using the "Teach" section of Washington, DC, Public Schools' (DCPS) IMPACT evaluation system. For the purposes of this study, any reference to the STAR3 observation tool refers to the "Teach" section of the IMPACT rubric used in the STAR3 project.

Typically, STAR3 Core teachers fall into one of two classifications: (1) elementary school teachers who teach state-tested subjects (reading, math, and 5th-grade science), or (2) middle school English Language Arts (ELA), math teachers, and 8th-grade science teachers. As part of the STAR3 program, all of these core teachers (N = 381) were observed one or two times during the 2011-12 school year using the STAR3 observation rubric (see appendix A). It should be noted that teachers of grades Kindergarten through 3rd grade were observed with the STAR3 instrument, as core teachers who received value-added scores based on student growth on the Iowa Test of Basic Skills. They are included in the sample of 381 described above, but as their value-added scores are not directly comparable to the EVAAS scores generated for teachers in grades 4-8, they were not included in the final 137 teachers used in this study.

The instrument itself features nine standards, with one item per standard. For each standard, the observer assigns a score on a four-point ordinal scale with labels: (1)

Ineffective, (2) Minimally Effective, (3) Effective, and (4) Highly Effective. The nine standards are:

Standard 1: Lead Well-Organized, Objective-Driven Lessons

Standard 2: Explain Content Clearly

Standard 3: Engage Students at All Learning Levels in Rigorous Work

Standard 4: Provide Students Multiple Ways to Engage with Content

Standard 5: Check for Student Understanding

Standard 6: Respond to Student Misunderstandings

Standard 7: Develop Higher-Level Understanding through Effective Questioning

Standard 8: Maximize Instructional Time

Standard 9: Build a Supportive, Learning-Focused Community

Teachers were observed by trained observers, all with classroom teaching experience. After a teacher's observation, observers provided both written and verbal feedback to the teacher, and provided copies of the written feedback to the teacher's STAR3 instructional coach for discussion. Individual teacher scores obtained on STAR3 observations were not provided to school administrators; only school-level aggregated data were given to administrators.

Observers were selected from an applicant pool of current and former educators and hired through the district HR process. The primary qualifications for hiring observers were a well-referenced tenure as a classroom teacher and experience observing and mentoring teachers. Training consisted of a two-week process in which observers

watched and discussed videos which were exemplars of the specific standard/rating combinations. These videos had been pre-scored by a panel of content experts. The expert three-person panel was selected by STAR3 and WS/FCS administrative personnel from program specialists in ELA and mathematics.

After inter-rater agreement on full-length lessons was consistently above 0.85 (for dichotomous categorization above/below "minimally effective"), observers conducted live practice observations in pairs in non-STAR3 schools. Overall inter-rater agreement was found to be 0.91 for dichotomous categorization above/below a score of 2.0), and observers were permitted to begin actual observations in STAR3 classrooms. All training and reliability calculations were performed by the STAR3 grant evaluator (this study's author). After the training was completed, actual observations began. In all, 625 observations were conducted on 381 teachers.

These observations took place during the 2011-12 school year. Approximately 4% (n = 23) of observations were conducted jointly by two observers for reliability purposes. For these observations, one of the observers was preselected as the "true" observer, and would conduct the observation in a typical fashion. The second, randomly selected observer would score the teacher as if the true observer, but did not provide written or oral feedback to the teacher. Scores on these reliability observations were collected and matched for the purposes of calculating inter-rater agreement; the results are shown in Table 6.

Table 6

Inter-rater Agreement of Paired STAR3 Observations

	Rater Agreement Percentage	
Standard	Exact	Classification
T1	78	91
T2	65	78
T3	70	91
T4	78	83
T5	70	87
T6	64	91
T7	55	77
T8	48	78
T9	78	78
Average	67	83

When scores were averaged across all nine standards (as is the STAR3 procedure for determining teacher effectiveness), exact rater agreement of the means occurred 22% of the time on the 1.0 – 4.0 scale, with a mean difference between paired raters of only 0.2. The primary purpose of this study was classifying teachers based on overall mean scores, so the classification agreement was calculated. Teachers scoring below 2.0 were classified as "low," teachers scoring between 2.0 and 2.99 were classified "middle," and teachers scoring at or above 3.0 were classified "high." Based on these classifications, paired observers agreed 19/23 times (83%).

The ultimate goal of this study was to examine the nature of the relationships among principals' evaluations of teachers (that are conducted based on formal and informal observations over the course of a full school year), third-party observations by trained observers without prior relationships with the teachers, and student outcomes.

Although the STAR3 instrument was designed to have nine independent standards, this assumption had not been empirically tested within this district.

An exploratory factor analysis of 2012-13 STAR3 observation data using principal components analysis led to the conclusion that a one-factor solution was the most reasonable. The one-factor model explained 46% of variance in scores (the second factor added another 10%), and the analysis produced only one eigenvalue with a value greater than one, as shown in the scree plot displayed in Figure 1. In addition, the Spearman correlation matrix of the nine standards with N = 866 observations showed significant correlations at the 0.01 level for all 36 pairs of standards ranging from 0.21 to 0.58, with only 5 of 36 correlations below 0.3. The one-factor solution was also shown to be a reliable measure, with an Alpha coefficient of 0.85. Assuming the validity of the one-factor model based on analysis of the 2012-13 data, the 2011-12 STAR3 observation data was averaged across standards to produce a single-score indicator to represent quality of teaching as measured by a trained observer.

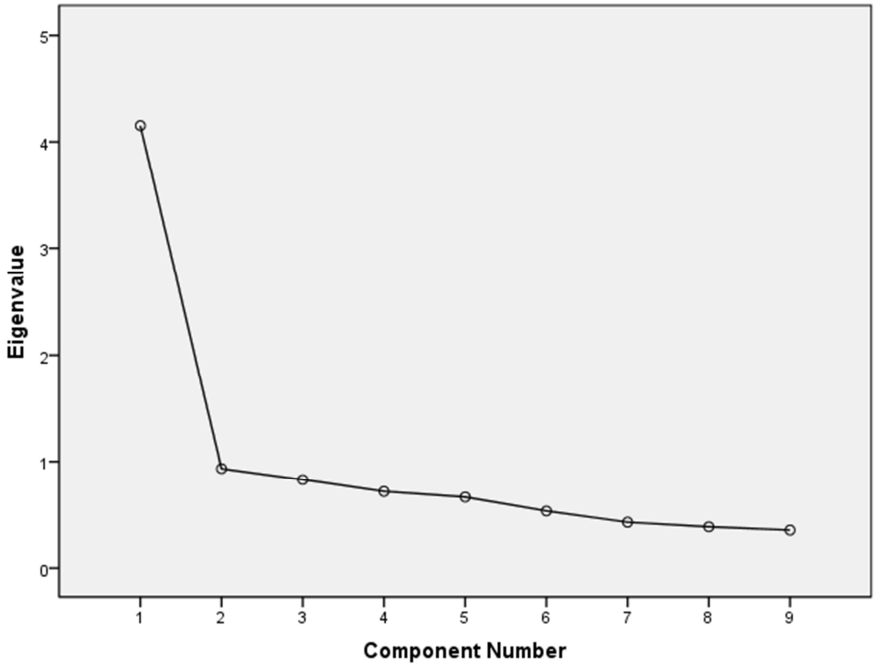

Figure 1. Scree Plot of 2012-13 STAR3 Observation Data

Data Analysis Procedures

RQ1: Do the methods of teacher effectiveness measurement classify teachers in substantively different ways?

To answer Research Question 1, multiple analyses were performed. First, correlation coefficients were calculated between each pair of the teacher scores on EVAAS® composite, STAR3 observation average, NC TES Standard I score, and NC TES Standard IV score. The strength and direction of these correlations provide insight into any potential redundancy in measurement among the methods. Scatterplots were also analyzed for the pairs of data, to determine if linear relationships existed.

As the focus of this study is primarily a comparison of the classification consistency among the various evaluation methods, the primary method of analysis for Research Question 1 was cross-tabular analysis. Each teacher was classified into a high-, middle- or low-achievement category on each of the measurements. The null hypothesis that the methods of classification were independent of one another was tested. In all, six tests were conducted – one for each of the six pairs of measurement methods. The NC TES scores are ordinal and categorical in nature, but as seen in Table 1, there were very few teachers classified at level two or below (developing or not demonstrated) on any of the standards. There were also very few teachers classified at level five (distinguished) on any standard. In practice, no reward is given for teachers whose TES scores are in a particular category, but teacher "action plans" are created for a teacher if his or her scores fall into the "developing" category or below. To reflect this practice, initial categorical analyses involving the TES considered teachers who scored in levels one or two to be "low," teachers in level three to be "middle," and teachers in levels four and five to be "high."

Like the TES scores, there was no reward for teachers who receive high ratings on the STAR3 instrument, but there is a potential penalty for very low scores. A teacher who averages below 2.0 on STAR3 observations receives a 25% reduction in any earned incentive pay. In practice, teachers receiving an average of 3.0 or higher on the STAR3 instrument are regarded as having taught highly effective lessons. Scores on individual standards on the STAR3 instrument range from 1-4 (discrete) and the overall score was calculated by taking the mean of the nine standards. For the STAR3 instrument, scores at

or above 1.0 but below 2.0 were classified "low," scores at or above 2.0 but below 3.0 were classified as "middle," and scores at or above 3.0 were considered "high." Mean STAR3 scores were treated as continuous variables.

In North Carolina, EVAAS® indices already divide teachers into three categories – below expected growth (less than -2.0), at expected growth (-2.0 to 2.0),and above expected growth (above 2.0). These designations were used as the low, middle, and high classifications for EVAAS®, respectively.

RQ2: In what ways are the current criteria for determining teacher pay in North Carolina (experience, advanced degrees, and National Board Certification), related to scores on different teacher effectiveness measures?

To answer Research Question 2, teachers' scores and classifications across measurement methods were examined with respect to pay variables. Of the three pay variables of interest, only one, years of teaching experience, was interval-level data. Correlations were calculated between teacher experience and each of the two measurement methods that were interval in nature, EVAAS composites and STAR3 observation scores.

To further examine the relationships between pay variables and classification via different measurement methods, cross tabulations and ANOVAs were the analyses used. A series of 2x3 cross-tabulation analyses were used to test the null hypothesis that a teacher's advanced degree status or National Board Certification status was independent of classification by each of the four methods.

To assess the relationship between teachers' years of experience and measurement classification, a series of ANOVAs was conducted with years of experience as the dependent variable and classification via each measurement method as the independent variable. These analyses tested the null hypothesis that there were no differences in mean teacher experience among levels of classification for each method of measurement.

RQ3: Which, if any, school-level variables have an impact on measures of teacher quality?

To address Research Question 3, first, two distributions were created. Each distribution ordered the sixteen project schools by either EVAAS® campus composite or STAR3 observation mean school score. EVAAS® campus composite was calculated like a teacher composite, but included all tests across all grades, rather than only those used for one individual teacher. In these analyses, NC TES scores were omitted because each school's NC TES scores are based on the ratings of their own principal. Inclusion of these scores would have resulted in difficulty separating actual school effects from principal rating effects.

Next, Spearman rank-order correlations were calculated to determine the degree of relationship between each of the measurement methods and the school variables. School variables included were school level (middle school/elementary school), school poverty (as measured by the percentage of students receiving free and reduced lunch), and the percentage of students who were English Language Learners. Whereas the strength and direction of these correlations was of interest in isolation, the primary

purpose of this analysis was to discover any differences in relationships between effectiveness measure and school variables. For example, one comparison of interest was to assess the strength of association between EVAAS® campus composite and school poverty versus the strength of association between STAR3 observation school average and school poverty. Significant differences here would suggest relative bias in one of the evaluation procedures.

Again, as the primary use of these teacher evaluation methods was for teacher classification, a 2x3 cross-tabulation analysis was used to test the null hypothesis that school level (a categorical variable) was independent of classification by each method. To test the relationship of school poverty and proportions of English language learners (interval variables) with measurement classification, a series of t-tests were conducted. The t-tests used free and reduced lunch percentage (FRL%) and English language learner percentage (ELL%) as the dependent variables and measurement classification as the independent variable to test the null hypotheses that: (1) There was no difference in FRL% or ELL% among classification levels grouped by EVAAS® composite, and (2) There was no difference in FRL% or ELL% among classification levels grouped by STAR3 observation score. Table 7 summarizes the three research questions, the data used in each, and the analyses conducted.

Table 7

Research Questions, Data Sources, and Analyses

	Data Source			
	EVAAS®	STAR3 observation	NC I	NC IV
RQ1: Do the methods of teacher effectiveness measurement classify teachers in substantively different ways?	X	X	X	X
RQ2: In what ways are the current criteria for determining teacher pay in North Carolina (experience, advanced degrees, and National Board Certification) related to scores on the different teacher effectiveness measures?	X	X	X	X
RQ3: Which, if any, school-level variables have an impact on measures of teacher quality?	X*	X*		

**Note: In RQ3, School average STAR3 score and EVAAS® Campus Composite are used*

	Analysis		
	Scatterplot/ Correlation	ANOVA/ t-test	Cross-tab/ Chi-Square
RQ1: Do the methods of teacher effectiveness measurement classify teachers in substantively different ways?	X		X
RQ2: In what ways are the current criteria for determining teacher pay in North Carolina (experience, advanced degrees, and National Board Certification) related to scores on the different teacher effectiveness measures?	X	X	X
RQ3: Which, if any, school-level variables have an impact on measures of teacher quality?	X	X	X

	Variables of Interest	
	Teacher-level	School-Level
RQ1: Do the methods of teacher effectiveness measurement classify teachers in substantively different ways?	X	
RQ2: In what ways are the current criteria for determining teacher pay in North Carolina (experience, advanced degrees, and National Board Certification) related to scores on the different teacher effectiveness measures?	X	
RQ3: Which, if any, school-level variables have an impact on measures of teacher quality?		X

CHAPTER IV

RESULTS

In this chapter, results from analyses performed to answer three research questions are described. Four methods of teacher effectiveness measurement were included: (1) a state-mandated principal rating of leadership (NC I), (2) a state-mandated principal rating of learning facilitation (NC IV), (3) a value-added measure of student growth (EVAAS), and (4) a third-party teacher observation score (STAR3 observation). In all, 137 teachers across 16 schools were included in the analyses. All schools had high proportions of students on free and reduced lunch and of English language learners. Teachers from grades 4-8 were included, as they were the only teachers in project schools for whom scores on all four measures existed.

Research Question 1: "Do the methods of teacher effectiveness measurement classify teachers in substantively different ways?"

To answer research question 1, teacher classification via each measurement method was analyzed. First, the distribution of each classification method was explored with descriptive statistics (see Table 8). STAR3 teacher observation scores are continuous with a possible range of 1.00 – 4.00. NC TES scores are categorical, with a possible minimum of 1 (Not Demonstrated) and a possible maximum of 5 (Distinguished). EVAAS® scores have no theoretical minimum or maximum; they are

index scores where scores below zero indicate performance below the state average and scores above zero indicate performance above the state average. Figures 2 and 3, and Table 9, show the frequencies of scores for each individual measurement.

Table 8

Descriptive Statistics for Each Measurement Method

Method	N	Minimum	Maximum	Mean	S.D.	Skew
EVAAS®	137	-10.11	3.03	-1.13	1.96	-1.02
NC I	137	2	5	3.43	.58	.74
NC IV	137	2	4	3.34	.50	.35
STAR3	137	1.63	3.89	2.89	.47	-.52

Table 9

Frequency Distributions of NC Standards I and IV for Sample

Standard	Not Demonstrated	Developing	Proficient	Accomplished	Distinguished
NC I	0	1	81	50	5
NC IV	0	2	87	48	0

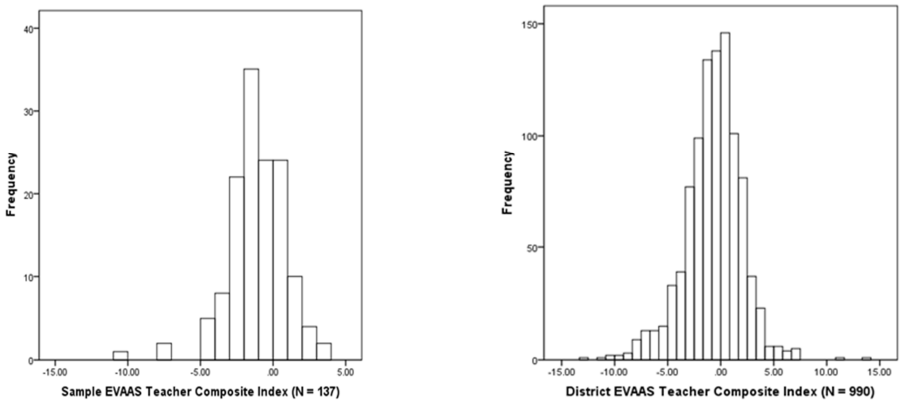

Figure 2. EVAAS® Teacher Composite Index Distribution for Sample and District

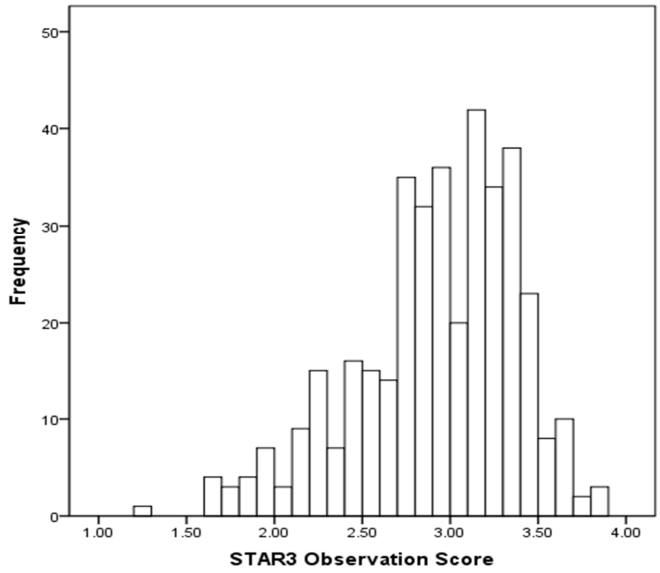

Figure 3. STAR3 Observation Score Distribution for Sample

When examining the EVAAS® data, three cases appeared to be outliers, as they lay below three standard deviations from the mean. The three values, -10.11, -7.66, and -7.21, lay 4.62, 3.36, and 3.13 standard deviations from the mean, respectively. However, the data for the entire district (not just STAR3 schools) represented a greater range. When

the STAR3 data were examined within the context of the EVAAS® scores for the entire district, these data all fell within 3.6 standard deviations of the mean. Additionally, removal of the three cases would only adjust the sample mean from -1.13 to -.98 – a difference of .15, or .08 of a standard deviation. For these reasons, in addition to having no reason to believe the scores were a result of error, the values were retained.

As discussed in Chapter 2, the scores on the NC TES tended to cluster in categories 3 and 4, with only a very small proportion of teachers receiving scores outside this range. In North Carolina, teachers are recommended for corrective action if a score of 2 or below is received. On both standards above, fewer than 1% of teachers received scores lower than 2. If one is to assume that 1% or more of teachers are performing below proficiency in reality, then, the NC TES is not adequately identifying these teachers. As shown in Figures 2 and 3, EVAAS® and STAR3 scores both showed greater variation than the NC TES standards, and both exhibited negative skew.

After reviewing the distributions created by each measurement method, a scatterplot or boxplot was created and Pearson correlation coefficients were calculated for each pair of measurement methods. These can be seen in Figures 4 through 9.

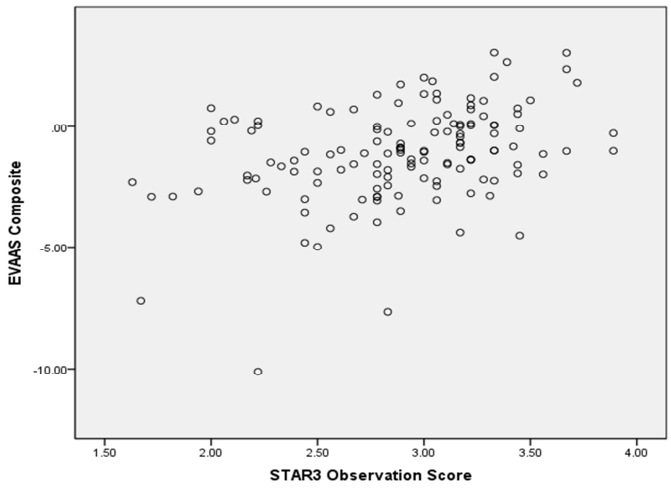

Figure 4. Scatterplot of EVAAS Index and STAR3 Observations (r = 0.36)**

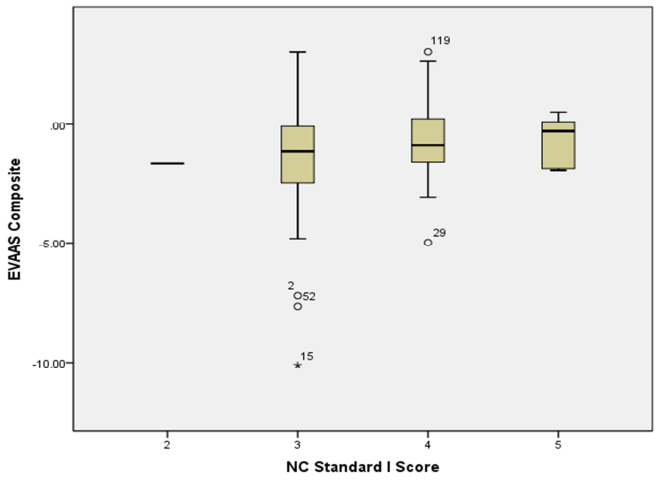

Figure 5. Boxplot of EVAAS Index and NC I (r = 0.19)*

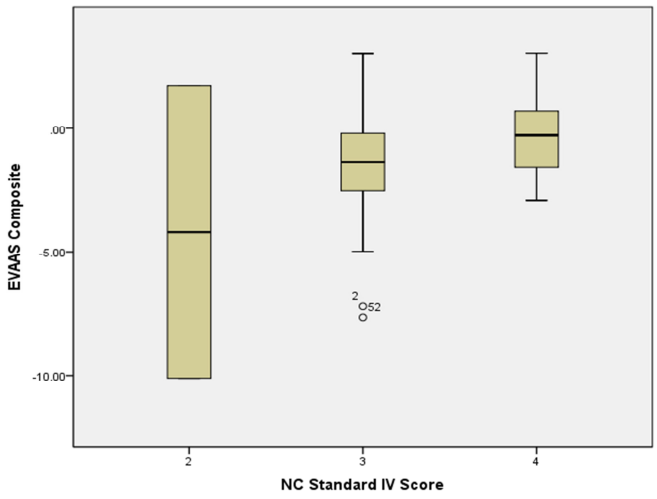

Figure 6. Boxplot of EVAAS Index and NC IV (r = 0.32)**

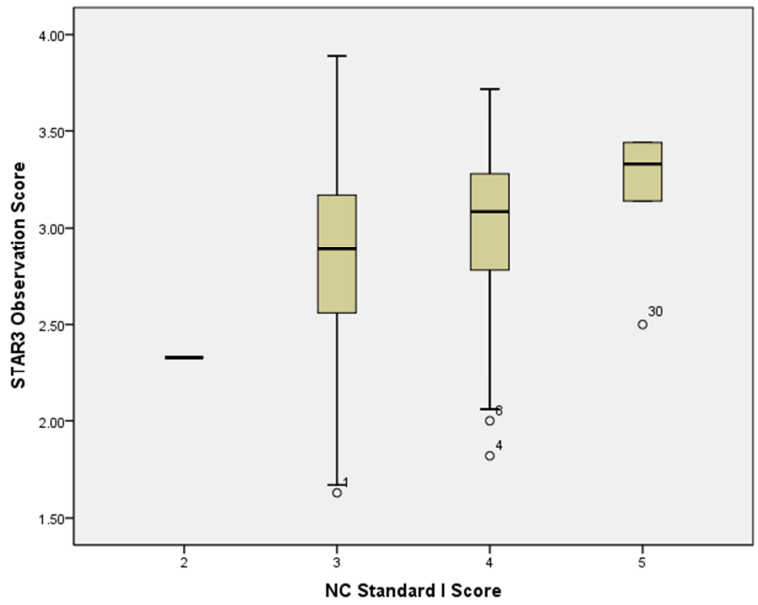

Figure 7. Boxplot of STAR3 Observations and NC I (r = 0.20)*

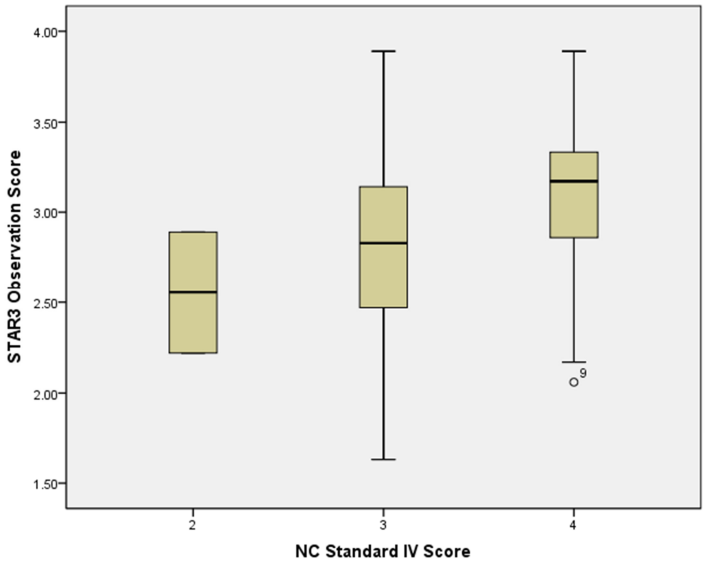

Figure 8. Boxplot of STAR3 Observations and NC IV (r = 0.33)**

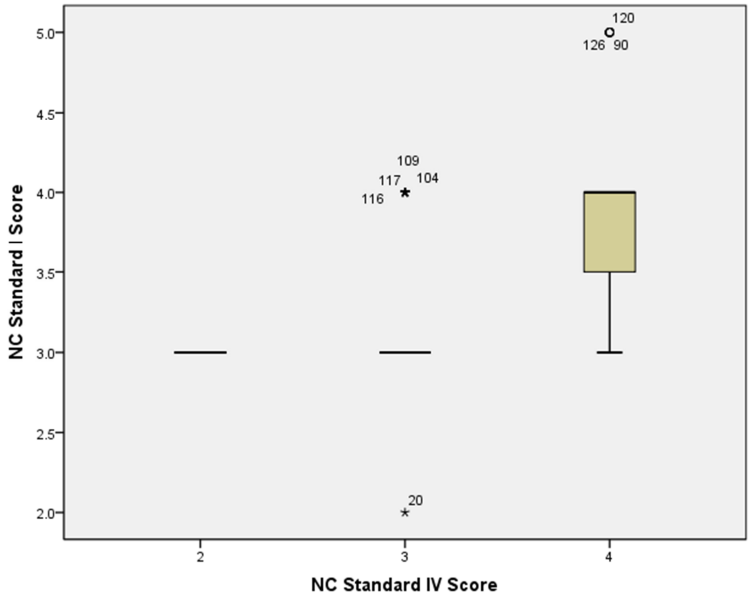

Figure 9. Boxplot of NC I and NC IV (r = 0.53)**

* Correlation is significant at the .05 level (2-tailed)

** Correlation is significant at the .01 level (2-tailed)

All correlations between pairs of measurement methods are statistically significant. Low-moderate, positive correlations were seen between STAR3 observations and EVAAS® (r = .36), STAR3 observations and NC IV (r = .33), and EVAAS® and NC IV (r = .32). These findings show that higher scores on NC IV were slighty associated with higher scores on both EVAAS® and STAR3 observations, and that higher scores on EVAAS® or STAR3 were slightly associated with higher scores on the other.

A moderate correlation existed between NC IV and NC I (r = .53), meaning that a higher score on NC I tended to be moderately associated with a higher score on NC IV. EVAAS® and NC I correlated at r = .19 and STAR3 observations and NC I correlated at r = .20, meaning that a higher score on NC I exhibited a weak association with higher scores on the non-NC TES measures.

To answer the research question, teachers were classified according to each measurement method as outlined in Chapter 3. As described in chapter 3, classification criteria for EVAAS® and NC TES are prescribed by the state of North Carolina. Teachers receiving EVAAS® indices below -2.0 are officially considered to be showing less than expected growth. Teachers with EVAAS® indices above 2.0 are officially classified as showing greater than expected growth. On the NC TES, scores of 1 or 2 are indicators of a teacher being “below proficiency,” and teachers are subject to personnel action based on this classification. There is no corresponding “high” category, but teachers with a score of 5 are considered “distinguished.” In the framework here, scores

of 1 or 2 were considered "low," scores of 3 were considered "middle," and scores of 4 or 5 were considered "high." For STAR3 observations, there is no such statewide classification. In practice in STAR3 schools, teachers receiving an observation average below 2.0 are subject to a reduction in incentive pay, so the "low" category in this project reflects that classification. Teachers scoring above 3.0 on the four-point scale have scores in the "highly effective" category, and the initial classification of "high" was set at this cut point. Table 10 shows the distribution of teachers in each classification by each measurement method.

Table 10

Classification Distributions by Measurement Method

	Classification		
Method	Low	Middle	High
EVAAS®	28%	68%	4%
NC I	1%	59%	40%
NC IV	1%	64%	35%
STAR3	4%	48%	48%

Except for EVAAS®, the measures rarely placed teachers in the "low" category. The STAR3 classification included very few teachers in the "low" classification, and relatively even amounts in the "middle" and "high" categories. For NC Standards I and IV, almost no teachers were classified as "low," the majority of teachers were placed in the "middle," and 35-40% placed in the "high" category.

Research question 1 asks if the measurement methods classify teachers in substantively different ways. To determine if there is a statistically significant difference

in classification, a series of cross tabular analyses were conducted, with the statistic of interest, when possible, being the Pearson Chi-square. For all of these tests, the null hypothesis being tested was that there is no association between the two methods of classification.

With four methods of classification (EVAAS®, STAR3, NC I, and NC IV), there were six pairs of methods to compare. With repeated statistical tests, it is wise to correct for a capitalization on chance when determining the rejection level (alpha) of a test statistic. For this project, a Bonferroni correction was applied. An overall alpha of .05 was desired, so with six comparisons, the alpha for each individual test was set at .008.

Tables 11-16 show the cross tabulations of each pair of measurement methods. Due to the low number of teachers categorized as "low" by both STAR3 and the NC TES measures and the low number of teachers categorized as "high" by EVAAS®, all tables contain 5/9 (55.6%) cells with an expected count below 5. Chi-square analyses are not considered appropriate in such conditions. Therefore, Cohen's Kappa was calculated for each analysis as a measure of agreement (see Table 17).

Table 11

STAR3 Classification vs. EVAAS® Classification

	EVAAS®			
STAR3	Low	Middle	High	Total
Low	5	0	0	5
Middle	23	43	0	66
High	10	50	6	66
Total	38	93	6	137

Table 12

NC I Classification vs. EVAAS® Classification

	EVAAS®			
NC I	Low	Middle	High	Total
Low	0	1	0	1
Middle	29	49	3	81
High	9	43	3	55
Total	38	93	6	137

Table 13

NC IV Classification vs. EVAAS® Classification

	EVAAS®			
NC IV	Low	Middle	High	Total
Low	1	1	0	2
Middle	32	52	3	87
High	5	40	3	48
Total	38	93	6	137

Table 14

NC I Classification vs. STAR3 Classification

	STAR3			
NC I	Low	Middle	High	Total
Low	0	1	0	1
Middle	4	44	33	81
High	1	21	33	55
Total	5	66	66	137

Table 15

NC IV Classification vs. STAR3 Classification

	STAR3			
NC IV	Low	Middle	High	Total
Low	0	2	0	2
Middle	5	47	35	87
High	0	17	31	48
Total	5	66	66	137

Table 16

NC IV Classification vs. NC I Classification

	NC I			
NC IV	Low	Middle	High	Total
Low	0	2	0	2
Middle	1	67	19	87
High	0	12	36	48
Total	1	81	55	137

Table 17

Agreement of Measurement Methods

Test	Cohen's Kappa	p
STAR3 - EVAAS®	0.06	0.17
NC I - EVAAS®	-0.07	0.09
NC IV - EVAAS®	-0.08	0.08
NC I - STAR3	0.16	0.04
NC IV - STAR3	0.18	0.02
NC I - NC IV	0.49	<0.001

As can be seen in Table 17, no pair of measurement methods exhibited statistically significant agreement, with the exception of the NC I – NC IV pair. Although the NC I – STAR3 and NC IV – STAR3 pairs exhibited a p-value below .05, a Bonferroni correction for six hypothesis tests yields a critical alpha of .008. From a policy perspective, substantial disagreement between methods of teacher effectiveness measurement could be problematic, particularly if there is a pattern of teachers scoring high on one measure while scoring low on another. The following are the measurement disagreements shown in tables 11-16 where these "major" disagreements (disagreement by more than one category) occured:

1. A teacher categorized as "low" by EVAAS® (n = 38) was twice as often categorized as "high" by STAR3 (n = 10) than as "low" (n = 5).

2. A teacher categorized as "high" by STAR3 (n = 66) was almost twice often categorized as "low" by EVAAS® (n = 10) than "high" (n = 6).

3. A teacher categorized as “high” by NC I (n =55) was three times as often categorized as “low” by EVAAS® (n = 9) than “high” (n = 3).

4. A teacher categorized as “low” by EVAAS® (n = 38) was more often categorized as “high” by NC IV (n = 5) than “low” (n = 1).

5. A teacher categorized as “high” by NC IV (n = 48) was more often categorized as “low” by EVAAS (n = 5) than “high” (n = 3).

These findings only address the disagreement between measurement methods when disagreement was by at least two categories (that is, when teachers were classified as both “low” and “high” by different measures). These types of major disagreements are the most problematic from a policy perspective.

As seen in Table 10, 48% or more of teachers are categorized into the “middle” by each measurement method. This high proportion of teachers being into the “middle” by all methods causes the highest marginal likelihood of any teacher, given any classification by any method, to be “middle,” with the following exceptions:

1. Teachers rated “low” by STAR3 were most often rated “low” by EVAAS.

2. Teachers rated “middle” by EVAAS were most often rated “high” by STAR3.

3. Teachers rated “high” by EVAAS were most often rated “high” by STAR3.

4. Teachers rated “high” by EVAAS were equally often rated “middle” or “high” by NC I.

5. Teachers rated “high” by EVAAS were equally often rated “middle” or “high” by NC IV.

6. Teachers rated “high” by STAR3 observation were equally often rated “middle” or “high” by NC I.

7. Teachers rated “high” by NC I were most often rated “high” by STAR3 observation.

8. Teachers rated “high” by NC IV were most often rated “high” by STAR3 observation.

9. Teachers rated “high” by NC IV were most often rated “high” by NC I

10. Teachers rated “high” by NC I were most often rated “high” by NC IV.

The measurement methods, when analyzed before teacher classification, exhibited significant small-to-moderate correlations. However, the analysis of classification agreement would suggest that the answer to research question 1 – whether the measurement methods classify teachers in substantively different ways – would be yes. There was substantive disagreement between methods in the classification of teachers.

Research Question 2: “In what ways are the current criteria for determining teacher pay in North Carolina (experience, advanced degrees, and National Board Certification) related to scores on the different teacher effectiveness measures?”

Although states take different approaches to aligning teacher effectiveness measurement and teacher compensation, North Carolina (like many other states), uses a simple formula when determining teacher pay. Teacher salaries, with some degree of district-by-district supplemental variation, are based on a teacher's years of experience, their attainment of degrees beyond the bachelor's, and whether or not they are certified by the National Board for Professional Teaching Standards (NBPTS). In Research Question 2, data were examined to determine if the four measurement methods categorize teachers differently based on teachers' values on the three pay-determining variables. In this sample, the vast majority of teachers were not National Board-certified. Approximately one-third of teachers held an advanced degree. There was a wide range of experience, with the average teacher having taught for 11.37 years (see Table 3 in Chapter 3).

Similarly to Research Question 1, correlations were calculated between each measurement method and the pay variables, with results shown in Table 18. Pearson product-moment correlations were calculated for pairs of continuous data and Spearman rank-order correlations were calculated for pairs with both continuous and categorical data.

Table 18

Correlations between Measurement Methods and Pay Variables

Variable Pair	Statistic	Value	p
EVAAS® - National Board	Spearman's ρ	-0.07	0.40
EVAAS® - Adv. Degree	Spearman's ρ	-0.07	0.44
EVAAS® - Experience	Pearson's r	0.10	0.25
NC I - National Board	Spearman's ρ	0.14	0.12
NC I - Adv. Degree	Spearman's ρ	0.36	<0.001
NC I - Experience	Spearman's ρ	0.26	<0.001
NC IV - National Board	Spearman's ρ	0.04	0.63
NC IV - Adv. Degree	Spearman's ρ	0.07	0.44
NC IV - Experience	Spearman's ρ	0.14	0.12
STAR3 - National Board	Spearman's ρ	-0.03	0.72
STAR3 - Adv. Degree	Spearman's ρ	0.04	0.65
STAR3 - Experience	Pearson's r	-0.05	0.54

No noteworthy correlations existed between measurement methods and pay variables, with the exception of a moderate, positive, statistically significant correlation of $\rho = .36$ (n = 137, $p < .001$) between NC TES Standard I and the presence of an advanced degree, and a small but statistically significant correlation ($r = .26$, n = 137, $p < .001$) between NC TES Standard I and teacher experience. These correlations show that, having an advanced degree tends to be moderately associated with a higher score on NC I and that having more years of teaching experience is also associated with a slightly higher score on NC I.

Since the research question concerns the effect of pay variables on classification, the next analyses performed were cross tabular analyses for the two categorical pay

variables. As with the cross tabular analyses in Research Question #1, all tables contained over 20% of cells with an expected value below 5, preventing a chi-square statistic to be calculated and interpreted with confidence. Tables 19 - 26 show the results for these analyses for degree type and NBPTS status.

Table 19

EVAAS® Classification by Degree Type

	EVAAS® Category			
Degree	Low	Middle	High	Total
Bach.	27	64	5	96
Adv.	11	29	1	41
Total	38	93	6	137

Table 20

EVAAS® Classification by NBPTS Status

	EVAAS® Category			
NB Status	Low	Middle	High	Total
No	35	90	5	130
Yes	3	3	1	7
Total	38	93	6	137

Table 21

NC Standard I Classification by Degree Type

	NC Standard I			
Degree	Low	Middle	High	Total
Bach.	1	67	28	96
Adv.	0	14	27	41
Total	1	81	55	137

Table 22

NC Standard I Classification by NBPTS Status

	NC Standard I			
NB Status	Low	Middle	High	Total
No	1	79	50	130
Yes	0	2	5	7
Total	1	81	55	137

Table 23

NC Standard IV Classification by Degree Type

	NC Standard IV			
Degree	Low	Middle	High	Total
Bach.	2	62	32	96
Adv.	0	25	16	41
Total	2	87	48	137

Table 24

NC Standard IV Classification by NBPTS Status

	NC Standard IV			
NB Status	Low	Middle	High	Total
No	2	83	45	130
Yes	0	74	3	7
Total	2	135	0	137

Table 25

STAR3 Classification by Degree Type

	STAR3 Score Category			
Degree	Low	Middle	High	Total
Bach.	3	49	44	96
Adv.	2	17	22	41
Total	5	66	66	137

Table 26

STAR3 Classification by NBPTS Status

	STAR3 Score Category			
NB Status	Low	Middle	High	Total
No	4	63	63	130
Yes	1	3	3	7
Total	13	167	201	137

Without a test for significance, the above analyses were limited to observing the proportions in each cell. It is possible that an association exists between NC Standard I and both National Board Status and Advanced Degree status, as over two-thirds (69.8%) of Bachelor's-level teachers were categorized as "middle" on NC I while almost the same proportion (65.9%) of advanced degree-holding teachers were categorized as "high."

Likewise, 71.4% of NB-certified teachers were classified as "high" compared to only 38.5% of non-NB-certified teachers. Among non-NB-certified teachers, the majority (60.8%) were categorized as "middle.) This finding agrees with the significant, moderate, positive correlation shown between advanced degree status and NC I, although no significant correlation was found between NB certification and NC I. No other pay variable-measurement method pairs appeared to exhibit a strong association.

To test the relationships of teacher experience (a continuous variable, unlike the other two pay variables) with the four classification methods, ANOVAs were performed with teacher experience as the dependent variable. See Tables 27-30 for the results of each ANOVA.

Table 27

ANOVA Results for Teacher Experience Means by EVAAS® Classification

Source	*df*	*F*	*p*	η^2	Power
EVAAS® Class	2	.39	.68	.01	.11
Error	134				
Total	136				

Table 28

ANOVA Results for Teacher Experience Means by NC I Classification

Source	*df*	*F*	*p*	η^2	Power
NC I Class	2	2.78	.07	.04	.54
Error	134				
Total	136				

Table 29

ANOVA Results for Teacher Experience Means by NC IV Classification

Source	*df*	*F*	*p*	η^2	Power
NC IV Class	2	.47	.62	.01	.13
Error	134				
Total	136				

Table 30

ANOVA Results for Teacher Experience Means by STAR3 Classification

Source	*df*	*F*	*p*	η^2	Power
STAR3 Class	2	2.16	.12	.03	.44
Error	134				
Total	136				

None of the four analyses was statistically significant. Based on the cross tabular and ANOVA results, there was little evidence to suggest that there is any substantial relationship between pay variables and classification on the four measurement methods in question. Therefore, Research Question 2, "in what ways are teacher pay variables related to scores on the teacher effectiveness measures," is best answered by stating that, aside from a possible relationship between NC TES Standard I (a measure of leadership) and possession of an advanced degree, there were no substantial relationships.

Research Question 3: "Which, if any, school-level variables have an impact on measures of teacher quality?"

To answer Research Question 3, the relationships between school-level demographic and effectiveness variables were examined. The EVAAS® Campus

Composite was used as the EVAAS® growth measure at the school level and each school's mean STAR3 teacher observation score was used for the school-level STAR3 observation score. The EVAAS® Campus Composite is based on the performance of all students in the school, rather than only students in one particular teacher's class. These two variables served as the measures of teacher effectiveness.

School-level demographic variables under consideration were school type (middle vs. elementary school), the percentage of students at the school who are English language learners (ELL%), and the percentage of students at the school who qualify for free or reduced-price lunch (FRL%). Table 31 shows descriptive statistics of each variable in the sample.

There were 16 schools, 12 of which were elementary and 4 of which were middle. All schools had a high (>80%) free and reduced lunch percentage. The range on English language learner percentage was much greater. EVAAS® campus composites had a wide range, though most were negative. A negative EVAAS® score implies performance below the state average, and a positive EVAAS® score implies performance above the state average.

The range of STAR3 school averages was smaller than the range of STAR3 teacher scores (see Research Question 1), which was to be expected.

Table 31

School-Level Variable Descriptive Statistics

Variable	Minimum	Maximum	Mean	SD
FRL%	80.97	99.19	92.88	5.05
ELL%	15.69	52.35	32.15	10.77
EVAAS®	-6.80	.40	-2.78	2.44
STAR3	2.62	3.20	2.90	.16

Before analyzing classification effects, correlations were calculated (see Table 32) to observe the degree of relationship between measurement methods, and between each measurement method and the school variables. Spearman's rank-order correlation was used with school type while Pearson's product-moment correlation was used for all other correlations.

Table 32

Correlations between Measurement Methods and School-Level Variables

Variable Pair	Statistic	Value	p
EVAAS® – School Type	Spearman's ρ	-0.33	0.21
EVAAS® – FRL%	Pearson's r	0.05	0.84
EVAAS® – ELL%	Pearson's r	-0.03	0.91
STAR3 – School Type	Spearman's ρ	-0.38	0.15
STAR3 – FRL%	Pearson's r	-0.23	0.39
STAR3 – ELL%	Pearson's r	0.29	0.28

No statistically significant correlations existed between measurement methods and school-level variables, though this may be due to the low number of schools in the analysis (N = 16). This indicates that there is not sufficient statistical evidence to assume a relationship between either STAR3 or EVAAS® campus composite and school type,

proportion of students receiving free or reduced lunch, or English language learner percentage. The low-moderate correlations observed between EVAAS® and school type (ρ = -.33, p = .21), and between STAR3 score and school type (ρ = -.38, p = .15), though not statistically significant, would provide the basis for an interesting repeat analysis with a larger sample, to see if a statistically significant association between elementary schools and higher scores on the STAR3 observation and EVAAS® exists.

To test for association between school type and the two methods of teacher evaluation, 2x3 cross tabular analyses were conducted (see Tables 33-34). Both EVAAS® and STAR3 observations were reviewed for association with school type.

Table 33

EVAAS® Classification by School Type

Type	Low	Middle	Total
Elem.	6	6	12
Middle	3	1	4
Total	9	7	16

Table 34

STAR3 Classification by School Type

Type	Middle	High	Total
Elem.	9	3	12
Middle	4	0	4
Total	13	3	16

In both cases, there did not appear to be a substantively different pattern between school type across levels of measurement categorization, and it was concluded that there was not sufficient evidence to show that classification via either measurement method is associated with whether the school in question is a middle school or an elementary school.

Given the other school-level variables in question (proportion of students receiving free or reduced lunch and English language learner percentage) were interval-level variables, and given the school-level teacher effectiveness classifications had only two levels, the formal test for determining a difference in proportion of students receiving free or reduced lunch or percentage of students who are English language learners across classifications by each measurement method was the t-test. Two t-tests were conducted (one each for FRL% and ELL%) for each measurement method. Table 35 shows the results of the t-tests comparing the FRL% group means for each classification level by each measurement method. It should be noted that FRL% was transformed using a square root transformation because its percentage values were all clustered above 80%. For the sake of interpretation, mean values for FRL% are reported in pre-transformation units.

Table 35

T-test Results for FRL% by Measurement Method

Method	Low Group Mean (SD)	Middle Group Mean(SD)	High Group Mean (SD)	*df*	*t*	*p*
EVAAS®	92.4 (5.2)	91.7 (5.3)	N/A	14	-0.33	0.747
STAR3	N/A	94.1 (5.1)	90.0 (2.9)	14	1.41	0.181

In the tests shown in Table 35, two null hypotheses were tested: (1) There was no significant mean difference in the proportion of students receiving free or reduced lunch among classification levels grouped by EVAAS® score; and (2) there was no significant mean difference in the proportion of students receiving free or reduced lunch among classification levels grouped by STAR3 observation scores. In both cases, the null hypotheses were not rejected, indicating that there was insufficient evidence to conclude that the proportion of students receiving free or reduced lunch and classification via either measurement method were related. However, it should be noted that achieved power for the t-tests was very low, at .06 for EVAAS and .30 for STAR3 scores. If significant differences existed, the small sample size would cause difficulty in detecting those differences. Even so, mean differences were small and therefore no practically significant differences were assumed.

Table 36 shows the results of the t-tests comparing the English language learner percentage group means for each classification level by each measurement method. In the results, two null hypotheses were tested: (1) there was no significant mean difference in English language learner percentage among classification levels grouped by EVAAS®

score; and (2) there was no significant mean difference in English language learner percentage among classification levels grouped by STAR3 observation score. In both cases, the null hypotheses were not rejected, indicating that there was insufficient evidence to conclude that English language learner percentage and classification via either measurement method are related. Again, however, observed power was low for these analyses, at .24 for EVAAS and .05 for STAR3. Such low power indicates that statistically significant differences, if present, would be difficult to identify, given the small sample size. Similar to the above t-tests, mean differences were small and therefore no practically significant differences were assumed.

Table 36

T-test Results for ELL% by Measurement Method

Method	Low Group Mean (SD)	Mid. Group Mean(SD)	High Group Mean (SD)	*df*	*t*	*p*
EVAAS®	35.2 (11.9)	28.3 (8.3)	N/A	14	1.29	0.218
STAR3	N/A	32.3 (12.0)	31.5 (3.1)	14	0.11	0.910

Admittedly, classification analyses at the school level are not perfect, as schools themselves are not classified by these means. These analyses are only intended to serve as an exploration into the relationships between measurement methods, classification, and certain school-level variables. In summary, there do not appear to be any relationships between measurement methods and school-level demographic variables, either through the raw EVAAS®/STAR3 scores or via classification. In light of these findings, Research

Question 3, “Which, if any, school-level variables have an impact on measures of teacher quality?” must be answered that no variables appear to have an impact.

CHAPTER V

DISCUSSION

This project examined teacher effectiveness data from 16 schools in a large urban district in North Carolina. The 16 schools were all participants in a Federal Teacher Incentive Fund grant program, and as such, were all high-poverty, low-achieving schools. Data collected from these teachers included a value-added measure of student growth, a classroom observation measure conducted by trained observers with a standardized rubric, and two measures required by the state to be scored by principals. Of the two principal ratings, one was a rating of teacher leadership and the other was a measure of pedagogy.

With much research currently being done on the effective use of teacher evaluation measures and their appropriate place in hiring, retaining, promoting, and firing teachers, this project aimed to contribute to the discussion by performing in-depth descriptive analyses on the effects of four distinct teacher evaluation measures. Of particular interest was the sorting of teachers into performance classes by each instrument and the variables that affect such groupings. If all measures were intended to assess and sort teachers into groups based on effectiveness, then differences in classification trends or differential effects of demographic variables would be noteworthy when determining how scores should be used. In these specific schools, the principal-rated measures are potentially used to make personnel actions at the low end of the scale, but are not used in

determining any positive interventions such as teacher pay. Conversely, the value-added measure of student growth is used to reward high-achieving teachers with incentive pay, while the classroom observation measure is used to reduce the incentive pay, if classification differences exist where the teacher earns high value-added scores but low observation scores. Observation results are also used as formative assessments for teacher development and growth.

The first research question answered in the project was, "Do the methods of teacher effectiveness measurement classify teachers in substantively different ways?" To answer this question, distributions were created and each teacher was classified into a "high," "middle," or "low" performance group by each measurement method. Correlations and cross tabulations were analyzed on each pair of measurement methods to determine if the methods classified teachers in meaningfully different ways.

Before analyzing the teacher classifications, the raw scores on each instrument were compared, and all pairs of measurement methods showed small, but statistically significant correlations ($r < .54$). Once classified, the only pair to show statistically significant agreement was the NC I – NC IV pair ($\kappa = 49$, $p < .001$). In other words, the principal ratings of leadership and pedagogy tended to agree on teacher classification. That would not be the case for any other pair of measurements. Based on prior research of principal ratings (e.g. Ho & Kane, 2013; Toch, 2008; Zatynski, 2012), principals have a tendency to rate teachers similarly across two standards.

When examining the classification of teachers on the value-added and observation methods more closely, every teacher classified as "high" by EVAAS® was also classified as such by STAR3, and every teacher classified as "low" by STAR3 was also classified as such by EVAAS®. However, there was a large amount of disagreement in classification between the two measures. For example, teachers classified as "low" by EVAAS® were twice as often categorized as "high" by STAR3 than as "low" by STAR3. Similarly, teachers categorized as "high" by STAR3 were almost twice as often categorized as "low" by EVAAS® than "high".

Among teachers categorized as "high" by NC I, three times as many teachers were categorized as "low" by EVAAS® than "high" by EVAAS®. Among teachers categorized as "low" by EVAAS®, more teachers were categorized as "high" by NC IV than as "low" by NC IV. Among teachers categorized as "high" by NC IV, teachers were more often categorized as "low" by EVAAS® than as "high" by EVAAS®. In essence, EVAAS® appears to have categorized teachers very differently than the other methods.

The second research question was, "In what ways are the current criteria for determining teacher pay in North Carolina (experience, advanced degrees, and National Board Certification) related to scores on the different teacher effectiveness measures?" It was examined in order to explore possible biases in one or more instruments toward teachers who differ on the variables that determine salary.

Correlational analyses showed no relationships between the pay variables and measurement methods, aside from a moderate ($r = 0.36$, $p < .001$) Spearman rank-order

correlation between NC Standard I (a principal rating of leadership in the school) and a teacher's having an advanced degree, and the same standard and teacher experience ($r = .26$, $p < .001$). Again, as the question refers to classification, ANOVAs and cross-tabular analyses were performed and no statistically significant relationships were found between pay variables and classification.

The third research question examined was, "Which, if any, school-level variables have an impact on teacher quality?" Currently, although teachers are, schools in North Carolina are not held to particular standards regarding value-added student growth or classroom observations of their teachers. However, under the current evaluation system, teachers for whom classroom-level growth data are not calculated are assigned the school EVAAS® composite as their individual growth measure. If school type, student poverty, or student language background were shown to impact school-level scores differentially across measurement methods, there could be implications for teacher recruitment, among other issues.

In the sample of 16 schools, there appeared to be no statistically significant relationships between school-level variables and school-level classification. Statistical power on these analyses was low due to the small sample size, but group means did not differ in any substantial way.

The most meaningful findings from the first research question were that all measurement methods were significantly, if modestly, positively correlated, but that only the two principal measures (NC I and NC IV) were shown to moderately agree on

classification. The fact that these measures of teacher effectiveness that collect vastly different types of data tended to be positively associated before classification speaks to the existence of a general teacher effectiveness value, and that the instruments were measuring it to some degree, although agreement among classifications does not reflect this. This finding also resembles that of Gersten et al. (2005) who found moderate correlations between observation data and student growth data for teachers.

However, despite the evidence for a general limited association between the measurement methods, there was much more disagreement than agreement when analyzing the categorization of teachers across the different instruments. In particular, the amount of disagreement across two categories (such as when a teacher is categorized as "high" on one measure and "low" on another, rather than "high" and "middle") is a disturbing finding if these results are to be used interchangeably. That is, if a district or state chooses to use only some of these four methods, valuable information about teacher quality may be left out. A teacher who is categorized as "high" by EVAAS® and "middle" by NC I or IV may very well be categorized as "low" by the STAR3 instrument, but the STAR3 instrument is not used in North Carolina schools outside of the STAR3 project. In fact, in this sample, teachers categorized as "high" by EVAAS® are more often rated "low" than "high" on STAR3. This suggests that either the two instruments are capturing very different dimensions of teaching quality, that one (or both) are measuring something unrelated to teacher quality, or that the methods are unreliable. Without evidence to suggest that the STAR3 observation is unreliable or invalid, adding the information provided by the STAR3 instrument actually provides a more complete

picture of the teacher's quality than EVAAS® and the NC TES alone. The STAR3 tool also has the added benefit of providing formative information to teachers, where EVAAS® does not (at least in terms of pedagogical practice).

The argument for adding STAR3 observation scores, or some other appropriate third-party observation measure to the measures currently in use in NC is stronger when one considers that formative feedback is a part of the STAR3 observation. In this sense, information is provided to help administrators identify effective and ineffective teachers, but teachers also benefit by gaining another, detailed perspective into the quality of their practice. However, as seen in the results, STAR3 score classification doesn't tend to agree highly with the other methods measured. A composite measurement system consisting of individual measures that disagree more often than they agree strains the credibility of the entire system. If none of these measures is universally recognized as the measure that most accurately captures what it means to be an effective teacher, then issues arise when associating these scores with rewards or personnel actions. Further investigation into the cause of this disagreement is recommended. One simple place to begin would be to adjust the classification cut scores (e.g., 3.5 being the minimum STAR3 observation score for "effective," rather than 3.0) and analyze the level of agreement. Since moderate correlational evidence for association between some of the methods exists, this adjustment may yield interesting findings. Additionally, it is problematic that there are no published reliability or validity statistics for the NC TES. Much research has been conducted on the validity and reliability of value-added teacher effectiveness scores, and some has been documented here. The inter-rater reliability of

the STAR3 observation tool has been shown to be high. The NC TES is the only measurement method analyzed in this project that has no published reliability or validity. With such high disagreement between measurement methods in this project, an investigation into the reliability of the NC TES might yield helpful results, particularly as the full NC TES currently comprises five of the six aspects of a teacher's official state evaluation.

The finding from the second research question that NC Standard I and advanced degrees are positively correlated is not a surprising one – it seems to suggest that teachers with advanced degrees tend to take on leadership roles within a school (or tend to be seen as leaders by principals). This finding on advanced degrees is in agreement with the research by Capur-Genturk et al. (2004) and Lease & Garrison (2008) that teachers with advanced degrees tend to score higher on some measures of teaching, if not on measures of student outcomes. Similarly, teachers with advanced degrees in this sample may take on leadership roles at a higher rate than their less-experienced peers, or be seen as doing so by principals.

The third research question examined the relationship between school-level variables and measurement methods. Principal evaluations of teachers were not included in these analyses, as each school's scores were assigned by its own principal and principal effect would be confounded with school effects.

It is a positive finding that there was no observed relationship between the measurement methods and either English language learner percentage or the proportion

of students receiving free or reduced lunch, as this suggests that student growth calculations and third-party observation scores for teachers are not impacted by a school's demographics in this particular sample.

The general EVAAS® model, as discussed in Ballou et al. (2004), does not include school-level covariates in its predictions, but does include a within-student covariance matrix. McCaffrey et al. (2004) show that the impact of school-level demographic variables are lessened when these matrices are used, and the findings in the present study would support the finding that demographic variables do not affect school-level composites when a within-student covariance matrix is used, as in EVAAS. There may be other reasons to contest the use of value-added modeling in schools, but this particular empirical application does not appear to exhibit demographic impact on scores. Again, it is worth noting that the sample is small and homogeneous, and potential relationships may be masked by insufficient statistical power, but mean differences across classifications were small.

In summary, the data analyzed in this project have shown that principal evaluations and STAR3 observations tend to classify teachers as moderate and effective much more often than ineffective, and that EVAAS® scores tend to classify teachers as ineffective and moderate much more often than effective (see Table 10 in Chapter 4). In general, all methods studied tend to disagree on teacher classification. The use of the STAR3 observation scores seems to be a valid addition to the evaluation system applied in STAR3 project schools, as it provides another level of usable data with regard to

identifying effective teachers. The STAR3 observations serve the added benefit of providing formative feedback for teachers who are concerned with professional growth – an element that is otherwise lacking. While NC TES measures may provide feedback to teachers, feedback from observers who are not teachers' direct supervisors can be beneficial in protecting against in-school bias (Bill & Melinda Gates Foundation, 2013).

Based on the findings of this small study, it appears that adding a third-party standards-based observation system to the NC TES would be helpful, as it serves the dual purpose of adding to the available data on teacher effectiveness and serving as a formative tool for teacher growth. In this sample, there is no evidence of demographic bias in EVAAS®, and so the retention of EVAAS® as a measure of teacher effectiveness in the NC TES is acceptable. However, the inclusion of a standards-based observation measure such as STAR3 would be crucial if the importance of using EVAAS® to determine personnel actions grows, since not all teachers who are exhibiting highly effective pedagogy are showing high levels of student growth. The two measures are related, but not redundant.

The incorporation of a third-party standards-based observation component to the state teacher evaluation system would also follow recommendations discussed by Darling-Hammond et al. (2012), Jacob & Lefgrin (2008), Staiger & Rockoff (2010), and Teddlie (2006).

As mentioned throughout the study, the sample size for this project was quite small and highly specific. All teachers sampled worked in high-poverty schools that had

been identified as high-need for the purposes of a federal grant intervention. It is very possible that findings from this study would not translate to a more diverse set of schools and teachers. Not only did the small sample size affect the data itself, in that lower-income schools and lower-growth student bodies were overrepresented relative to the state, but also in the statistical power of most analyses. Findings that may be true in the population would be difficult to perceive without a larger sample size. In addition, much of the data used violates assumptions of normality, further harming the generalizability of findings.

Another limitation of a study of this nature speaks to the difficulty in measuring teacher effectiveness at all. That is, "truth" is unknown. Comparing multiple measures of a single construct is difficult when none of the measures have been satisfactorily validated. When it is shown that STAR3 is more likely to identify teachers as highly effective than EVAAS®, for example, any conclusions are based purely on definition by the other instrument – hardly an ideal condition.

This particular project, because of its limitations, provides multiple directions for future research. The first and most obvious direction for a future study would be to replicate the analysis with a larger, more diverse sample – perhaps an entire district or state. Some of the findings of this project would carry a strong argument for policy revision if they were found in a sample more representative of the state at large. The structure of the study would remain intact. A larger number of schools randomly selected from the entire state of North Carolina could be sampled, and the same data – EVAAS®

indices, principal evaluations, and third-party standards-based observation scores (if available) could be examined. A larger, more representative sample would provide more diversity in growth scores and demographic variables, and more statistical power for analyses.

Another study of interest would be to analyze teachers' perceptions of the relative value and accuracy of the four measurement methods studied. As Papay (2012) notes, effective teacher evaluation serves two purposes: to identify effective and ineffective teachers, and to provide formative feedback for teacher growth. STAR3 project teachers could be interviewed and surveyed to determine their opinions on the fairness and accuracy of the methods (a measure of the power of identification) and their opinions on the usefulness of results from each method (a measure of the formative value).

As Koedel and Betts (2007) and McCaffrey et al. (2009) found, year-to-year correlations of value-added teacher effect tend to be in the 0.2 to 0.5 range. A longitudinal study of this particular sample analyzing the trend of group categorization via EVAAS®, as well as the other teacher effectiveness measures, would provide valuable insight into the comparative mobility from group-to-group experienced among teachers across measures. If particular methods were shown to provide more stable results year to year, then those methods may have additional value in assessing a teacher's persistent quality.

Finally, the finding that the most experienced teachers tended to score higher on all measures of teacher effectiveness other than STAR3 provides room for exploration. A

qualitative analysis of the feedback given to teachers across experience levels would be beneficial, as would a focused study, involving teachers being observed using multiple rubrics, with each rubric containing different measures of pedagogical skill. Findings from this study could inform the use of standards-based observations in state teacher evaluation systems.

With the availability of data on teacher effectiveness growing annually, the emphasis on identifying high-quality teachers will continue to grow. Stringent analysis of the instruments and methods used in these processes will serve to ensure that teacher evaluation remains fair, efficient, and beneficial.

REFERENCES

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational researcher*, *37*(2), 65-75.

Anderson, J. (2013). Curious grade for teachers: Nearly all pass. *The New York Times*. Retrieved from http://www.nytimes.com/2013/03/31/education/curious-grade-for-teachers-nearly-all-pass.html?pagewanted=all.

Arrington, S. E. (2010). *Elementary principals' follow-through in teacher evaluation to improve instruction.* (Doctoral Dissertation). Georgia Southern University, Statesboro, GA.

Ballou, D. (2002). Sizing up test scores. *Education Next*, *2(2),* 10-15.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29(1),* 37-65.

Banchero, S. & Kesmodo, D. (2011). Teachers are put to the test. *The Wall Street Journal*. Retrieved from http://online.wsj.com/news/articles/SB10001424053111903895904576544523666

Bill and Melinda Gates Foundation. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Policy and Practice Summary. *MET Project*.

Bill and Melinda Gates Foundation. (2013). Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study. Policy and Practice Brief. *MET Project*.

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics*, *36(5),* 616-637.

Collins, C., and Amrein-Beardsley, A. (2013). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116(1).*

Copur-Gencturk, Y., Hug, B., & Lubienski, S. T. (2014). The effects of a master's program on teachers' science instruction: Results from classroom observations, teacher reports, and student surveys. *Journal of Research in Science Teaching*, *51(2)*, 219-249.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93(6),* 8-15.

Gersten, R., Baker, S. K., Haager, D., & Graves, A. W. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners: An observational study. *Remedial & Special Education, 2(4),* 197-206.

Gitomer, D. H. (2011). Road maps for learning and teacher evaluation. *Measurement: Interdisciplinary Research and Perspectives, 9,* 146-148.

Goldhaber, D., Perry, D., & Anthony, E. (2004). The National Board for Professional Teaching Standards (NBPTS) process: Who applies and what factors are associated with NBPTS certification? *Educational Evaluation and Policy Analysis, 26(4)*, 259-280.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education*, *4(4),* 319-350.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48(3),* 794-831.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41(2)*, 56-64.

Ho, A. D., Kane, T. J., & Bill and Melinda Gates Foundation. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. *MET Project*. Ingle, K., Rutledge, S., & Bishop, J. (2011). Context matters: principals' sensemaking of teacher hiring and on-the-job performance. *Journal of Educational Administration, 49(5)*, 579-610.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26(1),* 101-136.

Johnson, B. L. J. (1997). An organizational analysis of multiple perspectives of effective teaching: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11(1)*, 69-87.

Johnson, S. D., & Roellke, C. F. (1999). Secondary teachers' and undergraduate education faculty members' perceptions of teaching-effectiveness criteria: A national survey. *Communication Education, 48(2)*, 127-138.

Keeves, J. P., Hungi, N., & Afrassa, T. (2005). Measuring value added effects across schools: Should schools be compared in performance? *Studies in Educational Evaluation*, *31*(*2)*, 247-266.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. National Center on Performance Incentives, Vanderbilt, Peabody College.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, *25(3)*, 287-298.

Ladd, H. & Walsh, R. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review, 21(1),* 1-17.

Lease, A., & Garrison, L. F. (2008). Improving teacher effectiveness through focused graduate education. *Action in Teacher Education, 30(3)*, 11-22.

Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis*, *34(1),* 109-121.

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44(1),* 47-67.

Marder, M. (2012). Measuring teacher quality with value-added modeling. *Kappa Delta Pi Record, 48(4),* 156-161.

Marshall, K. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan*, *86(10),* 727-735.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31(1),* 35-62.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29(1),* 67-101.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education*, *4(4),* 572-606.

McConney, A., Ayres, R., Hansen, J. B., & Cuthbertson, L. (2003). Quest for quality: recruitment, retention, professional development, and performance evaluation of teachers and principals in Baltimore City's public schools. *Journal of Education for Students Placed at Risk, 8(1)*, 87-116.

McREL. (2009). *North Carolina Teacher Evaluation Process*. Teacher Evaluation Instrument. Retrieved from http://www.ncpublicschools.org/docs/effectiveness-model/ncees/instruments/teach-eval-manual.pdf

National Council on Teacher Quality, (2013). *State of the states 2012: Teacher effectiveness policies*. Washington, DC:

Noakes, L. A. (2008). Adapting the utilization-focused approach for teacher evaluation. *Journal of Multidisciplinary Evaluation*, *6(11)*, 83-88.

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82(1),* 123-141.

Penny, J. M. & Ward, M. (2012, February). *Predictors of teacher effectiveness using EVAAS value-added scores.* Paper presented at the Annual Meeting of the North Carolina Association for Research in Education, Winston-Salem, NC.

Ramirez, A., Lamphere, M., Smith, J., Brown, S., & Pierceall-Herman, J. (2011). Teacher development and evaluation: A study of policy and practice in Colorado. *Management in Education, 25*(3), 95-99.

Rice, J. K. (2010). The impact of teacher experience: Examining the evidence and policy implications. Brief No. 11. *National Center for Analysis of Longitudinal Data in Education Research*.

Ripley, A. (2010, January 1). What Makes a Great Teacher? *The Atlantic*. Retrieved from http://www.theatlantic.com/magazine/archive/2010/01/what-makes-a-great-teacher/307841/

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(*1)*, 175-214.

Rothstein, J. (2012). *Teacher quality policy when supply matters*. (No. w18419). National Bureau of Economic Research.

Rouse, W, & Holloman, H. (2005). A comparison of student test results: Business and marketing education National Board certified teachers and non-National Board certified teachers. *Delta Pi Epsilon Journal*, *47(3),*128-142.

Sanders, W. (2006, October). Comparisons among various educational assessment value-added models. In *National Conference on Value-Added*.

Sanders, W., Saxton, A. & Horn, S. (1997). The Tennessee value-added assessment system. In Millman, J. (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Santelices, M. V., & Taut, S. (2011). Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system. *Assessment in Education: Principles, Policy & Practice, 18(1),* 73-93.

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin*, *95(2),* 122-140.

Searles, W. E., & Ng, R. W. M. (1982). A comparison of teacher and principal perception of an outstanding biology teacher. *Journal of Research in Science Teaching, 19(6)*, 487-95.

Searles, W. E., & Kudeki, N. (1987). A comparison of teacher and principal perception of an outstanding science teacher. *Journal of Research in Science Teaching, 24(1)*, 1-13.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *The Journal of Economic Perspectives*, *24*(*3)*, 97-117.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62(4)*, 339-355.

Stronge, J., Ward, T., Tucker, P., Hindman, J., McColsky, W., & Howard, B. (2007). National Board certified teachers and non-National Board certified teachers: Is there a difference in teacher effectiveness and student Achievement? *Journal of Personnel Evaluation in Education, 20,* 3-4.

Teacher Incentive Fund. (2010). Summary of notice of proposed priorities. Retrieved from http://www2.ed.gov/programs/teacherincentive/summarymemo32010.doc

Teddlie, C., Creemers, B., Kyriakides, L., Muijs, D., & Yu, F. (2006). The international system for teacher observation and feedback: Evolution of an international study of teacher effectiveness constructs. *Educational Research and Evaluation, 12(6)*, 561-582.

Timmermans, A. C., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement*, *22(4)*, 393-413.

Toch, T. (2008). Fixing teacher evaluation. *Educational Leadership, 66(2),* 32-37.

Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School effectiveness and school improvement*, *20(2)*, 269-285.

Van Tassel-Baska, J., Quek, C., & Feng, A. X. (2006). The development and use of a structured teacher observation scale to assess differentiated best practice. *Roeper Review, 29(2),* 84-92.

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). SAS® EVAAS® statistical models. *SAS White Paper*.

Zatynski, M. (2012). Revamping teacher evaluation. *Principal, 91(5),* 22-27.

Zimmerman, S., & Deckert-Pelton, M. (2003). Evaluating the evaluators: Teachers' perceptions of the principal's role in professional evaluation. *NASSP Bulletin, 87(636),* 28-37.

APPENDIX A

STAR3 OBSERVATION RUBRIC

T1: Lead Well-organized, Objective-Driven Lessons:

Highly Effective (4)

• Students can authentically explain *what* they are learning, beyond simply repeating back the stated or posted objective.

• Students can authentically explain *why* what they are learning is important, beyond simply repeating the teacher's explanation.

• Students understand how the objective fits into the broader unit and course goals. For example, this might be shown through an effective teacher explanation of how the lesson connects to the unit's essential questions or structure, or reflected in students demonstrating through their comments that they understand how the lesson fits into the broader goals of the unit.

• The teacher actively and effectively engages students in the process of connecting the lesson to their prior knowledge. For example, the teacher might ask students to connect concepts to their own experiences or to what they have learned in other courses.

Effective (3)

• The lesson objective is specific, measurable, and aligned to standards; it conveys what students are learning and what they will be able to do by the end of the lesson.

• The objective of the lesson is clear to students. For example, the teacher might clearly state and explain the objective, or students might demonstrate through their actions that they understand what they will be learning and doing.

• The teacher ensures that students understand the importance of the objective. For example, the teacher might effectively explain its importance, or students might demonstrate through their comments that they understand the importance of what they are learning.

• The lesson builds on students' prior knowledge in a significant and meaningful way, as appropriate to the objective.

• The lesson is well-organized: All parts of the lesson are connected to each other and aligned to the objective, and each part significantly moves students toward mastery of the objective.

Minimally Effective (2)

• The lesson objective may be missing one component (for example, it might not be specific, or it might not be aligned to standards), but it does convey what students are learning and what they will be able to do by the end of the lesson.

• The teacher may state the objective of the lesson but may do so in a way that does not effectively lead to student understanding. For example, the objective might not be in developmentally appropriate language.

• The teacher may explain the importance of the objective but may do so in a way that does not effectively lead to student understanding. For example, the explanation might be too general to be effective.

• The teacher may state how the lesson connects to students' prior knowledge, but the lesson generally does not build on students' prior knowledge in a significant and meaningful way. For example, the teacher might simply make a reference to what students were doing in the previous lesson.

• Some parts of the lesson may not be closely connected to each other or aligned to the objective, or some parts may not significantly move students toward mastery of the objective.

Ineffective (1)

• The lesson objective may be missing more than one component, the objective may not convey what students are learning or what they will be able to do by the end of the lesson, there may not be a clear
objective to the lesson, or the objective stated or posted may not connect to the lesson taught.

• The teacher may not state the objective, or students may be unclear or confused about what they will be learning and doing.

• The teacher may not explain the importance of the objective, or students may not understand its importance.

• The teacher may make no effort to have the lesson build on or connect to students' prior knowledge, or the teacher may make an effort that is ineffective.

• The lesson may be generally disorganized. Different parts of the lesson may have no connection to each other, students may be confused about what to do, most parts of the lesson may not be aligned to the objective, or most parts of the lesson may not significantly move students toward mastery of the objective.

T2: Explain Content Clearly

Highly Effective (4)

• Explanations are concise, fully explaining concepts in as direct and efficient a manner as possible.

• The teacher effectively makes connections with other content areas, students' experiences and interests, or current events in order to make the content relevant and build student understanding and interest.

• When appropriate, the teacher explains concepts in a way that actively involves students in the learning process, such as by facilitating opportunities for students to explain concepts to each other.

• Explanations provoke student interest in and excitement about the content.

• Students ask higher-order questions and make connections independently, demonstrating that they understand the content at a higher level.

Effective (3)

• Explanations of content are clear and coherent, and they build student understanding of content.

• The teacher uses developmentally appropriate language and explanations.
• The teacher gives clear, precise definitions and uses specific academic language as appropriate.

• The teacher emphasizes key points when necessary.

• When an explanation is not effectively leading students to understand the content, the teacher adjusts quickly and uses an alternative way to effectively explain the concept.

• Students ask relatively few clarifying questions because they understand the explanations. However, they may ask a number of extension questions because they are engaged in the content and eager to learn more about it.

Minimally Effective (2)

• Explanations are generally clear and coherent, with a few exceptions, but they may not be entirely effective in building student understanding of content.

• Some language and explanations may not be developmentally appropriate.

• The teacher may sometimes give definitions that are not completely clear or precise, or sometimes may not use
academic language when it is appropriate to do so.

• The teacher may only sometimes emphasize key points when necessary, so that students are sometimes unclear
about the main ideas of the content.

• When an explanation is not effectively leading students to understand the concept, the teacher may sometimes
move on or re-explain in the same way rather than provide an effective alternative explanation.

• Students may ask some clarifying questions showing that they are confused by the explanations.

Ineffective (1)

• Explanations may be unclear or incoherent, and they are generally ineffective in building student understanding of content.

• Much of the teacher's language may not be developmentally appropriate.

• The teacher may frequently give unclear or imprecise definitions, or frequently may not use academic language when it is appropriate to do so.

• The teacher may rarely or never emphasize key points when necessary, such that students are often unclear about the main ideas of the content.

• The teacher may frequently adhere rigidly to the initial plan for explaining content even when it is clear that an explanation is not effectively leading students to understand the concept.

• Students may frequently ask clarifying questions showing that they are confused by the explanations, or students may be consistently frustrated or disengaged because of unclear explanations.

T3: Engage Students at All Learning Levels in Rigorous Work

Highly Effective (4)

• The teacher makes the lesson accessible to all students at different learning levels.

• The teacher makes the lesson challenging to all students at different learning levels.

Effective (3)

• The teacher makes the lesson accessible to almost all students; there is evidence that the teacher knows each student's level and ensures that the lesson meets almost all students where they are. For example, if necessary, the teacher might differentiate content, process, or product (using strategies that might include, for example, flexible grouping, leveled texts, or tiered assignments) in order to ensure that students are able to access the lesson.

• The teacher makes the lesson challenging to almost all students; there is evidence that the teacher knows each student's level and ensures that the lesson pushes almost all students forward from where they are. For example, the teacher might ask more challenging questions, assign more demanding work, or provide extension assignments in order to ensure that all students are challenged by the lesson.

• There is an appropriate balance between teacher-directed instruction and rigorous student-centered learning during the lesson, such that students have adequate opportunities to meaningfully practice, apply, and demonstrate what they are learning.

Minimally Effective (2)

• The teacher makes the lesson accessible to most students; some students may not be able to access certain parts of the lesson.

• The teacher makes the lesson challenging to most students; some students may not be challenged by certain parts of the lesson.

• While students have some opportunities to meaningfully practice, apply, and demonstrate what they are learning, there is more teacher-directed instruction than appropriate.

Ineffective (1)

•The lesson is not accessible to most students.

• The lesson is not challenging to most students.

• The lesson is almost entirely teacher-directed, and students have few opportunities to meaningfully practice, apply, and demonstrate what they are learning.

T4: Provide Students Multiple Ways to Engage With Content

Highly Effective (4)

• The ways students are provided to engage with content all significantly promote student mastery of the objective; students respond positively and are actively involved in the work.

Effective (3)

• The teacher provides students more than one way to engage with content, as appropriate, and all ways are matched to the lesson objective. For particular types of lessons, this may only entail giving students two ways to engage with content (for example, a Socratic seminar might involve verbal/linguistic and interpersonal ways), while for many lessons, this may involve three or more.

• The ways students engage with content all promote student mastery of the objective.

Minimally Effective (2)

• The teacher provides students more than one way to engage with content, but not all of these may be well matched to the lesson objective; or, the teacher may only give students two ways to engage with content when using an additional way would have been more appropriate to the objective (for example, a lesson introducing fractions that involves only auditory and interpersonal but not visual or tactile/kinesthetic ways).

• Some ways provided do not promote student mastery of the objective.

Ineffective (1)

• The teacher provides students with more than one way to engage with content, but most of these may not be well matched to the lesson objective; or, the teacher may only give students one way to engage with the content.

• Most or all ways provided do not promote student mastery of the objective; or, some ways may detract from or impede student mastery.

T5: Check For Student Understanding

Highly Effective (4)

• The teacher checks for understanding at all key moments.

• Every check gets an accurate "pulse" of the class's understanding.

• The teacher uses a variety of methods of checking for understanding.

• The teacher seamlessly integrates information gained from the checks by making adjustments to the content or delivery of the lesson, as appropriate.

Effective (3)

• The teacher checks for understanding of content at almost all key moments (when checking is necessary to inform instruction going forward, such as before moving on to the next step of the lesson or partway through the independent practice).

• The teacher gets an accurate "pulse" of the class's understanding from almost every check, such that the teacher has enough information to adjust subsequent instruction if necessary.

• If a check reveals a need to make a whole-class adjustment to the lesson plan (for example, because most of the students did not understand a concept just taught), the teacher makes the appropriate adjustment in an effective way.

Minimally Effective (2)

• The teacher sometimes checks for understanding of content, but misses several key moments.

• The teacher gets an accurate "pulse" of the class's understanding from most checks.

• If a check reveals a need to make a whole-class adjustment to the lesson plan, the teacher attempts to make
the appropriate adjustment but may not do so in an effective way.

Ineffective (1)

• The teacher rarely or never checks for understanding of content, or misses nearly all key moments.

• The teacher does not get an accurate "pulse" of the class's understanding from most checks. For example, the teacher might neglect some students or ask very general questions that do not effectively assess student understanding.

• If a check reveals a need to make a whole-class adjustment to the lesson plan, the teacher does not attempt to make the appropriate adjustment, or attempts to make the adjustment but does not do so in an effective way.

T6: Respond To Student Misunderstandings

Highly Effective (4)

• The teacher responds to almost all student misunderstandings with effective scaffolding.

• The teacher anticipates student misunderstandings and preemptively addresses them, either directly or through the design of the lesson.

• The teacher is able to address student misunderstandings effectively without taking away from the flow of the lesson or losing the engagement of students who do understand.

Effective (3)

• The teacher responds to most student misunderstandings with effective scaffolding.

• When possible, the teacher uses scaffolding techniques that enable students to construct their own understandings (for example, by asking leading questions) rather than simply re-explaining a concept.

• If an attempt to address a misunderstanding is not succeeding, the teacher, when appropriate, responds with another way of scaffolding.

Minimally Effective (2)

•The teacher responds to some student misunderstandings with effective scaffolding.

• The teacher may primarily respond to misunderstandings by using scaffolding techniques that are teacher-driven (for example, re-explaining a concept) when student-driven techniques could have been effective.

• The teacher may sometimes persist in using a particular technique for responding to a misunderstanding, even when it is not succeeding.

Ineffective (1)

• The teacher responds to few student misunderstandings with effective scaffolding.

• The teacher may only respond to misunderstandings by using scaffolding techniques that are teacher-driven when student-driven techniques could have been effective.

• The teacher may frequently persist in using a particular technique for responding to a misunderstanding, even when it is not succeeding.

T7: Develop Higher-level Understanding Through Effective Questioning

Highly Effective (4)

• The teacher asks higher-level questions at multiple levels of Bloom's taxonomy, if appropriate to the lesson.

• Students are able to answer higher-level questions with meaningful responses, showing that they are accustomed to being asked these kinds of questions.

• Students pose higher-level questions to the teacher and to each other, showing that they are accustomed to asking these questions.

Effective (3)

• The teacher frequently develops higher-level understanding through effective questioning.

• Nearly all of the questions used are effective in developing higher-level understanding.

• The teacher uses a variety of questions.

Minimally Effective (2)

• The teacher sometimes develops higher-level understanding through effective questioning.

• Some of the questions used may not be effective in developing higher-level understanding. For example, the teacher might ask questions that are unnecessarily complex or confusing to students.

• The teacher may repeatedly use two or three questions.

Ineffective (1)

• The teacher rarely or never develops higher-level understanding through effective questioning.

• Most of the questions used may not be effective in developing higher-level understanding. For example, the teacher might ask questions that do not push students' thinking.

• The teacher may only use one question repeatedly. For example, the teacher might always ask students "Why?" in response to their answers.

T8: Maximize Instructional Time

Highly Effective (4)

• Routines and procedures run smoothly with minimal prompting from the teacher; students know their responsibilities and do not have to ask questions about what to do.

• Transitions are orderly, efficient, and systematic, and require little teacher direction.

• Students are never idle while waiting for the teacher (for example, while the teacher takes attendance or prepares materials).

• Students share responsibility for the operations and routines in the classroom.

• The lesson progresses at a rapid pace such that students are never disengaged, and students who finish assigned work early have something else meaningful to do.

• The flow of the lesson is never impeded by inappropriate or off-task student behavior, either because no such behavior occurs or because when such behavior occurs the teacher efficiently addresses it.

Effective (3)

• Routines and procedures run smoothly with some prompting from the teacher; students generally know their responsibilities.

• Transitions are generally smooth with some teacher direction.

• Students are only idle for very brief periods of time while waiting for the teacher (for example, while the teacher takes attendance or prepares materials).

• The teacher spends an appropriate amount of time on each part of the lesson.

• The lesson progresses at a quick pace, such that students are almost never disengaged or left with nothing meaningful to do (for example, after finishing the assigned work, or while waiting for one student to complete a problem in front of the class).

• Inappropriate or off-task student behavior rarely interrupts or delays the lesson.

Minimally Effective (2)

• Routines and procedures are in place but require significant teacher prompting and direction; students may be unclear about what they should be doing and may ask questions frequently.

• Transitions are fully directed by the teacher and may be less orderly and efficient.

• Students may be idle for short periods of time while waiting for the teacher.

• The teacher may spend too much time on one part of the lesson (for example, may allow the opening to continue longer than necessary).

• The lesson progresses at a moderate pace, but students are sometimes disengaged or left with nothing meaningful to do.

• Inappropriate or off-task student behavior sometimes interrupts or delays the lesson.

Ineffective (1)

• There are no evident routines and procedures, so the teacher directs every activity; students are unclear about what they should be doing and ask questions constantly or do not follow teacher directions.

• Transitions are disorderly and inefficient.

• Students may be idle for significant periods of time while waiting for the teacher.

• The teacher may spend an inappropriate amount of time on one or more parts of the lesson (for example, spends 20 minutes on the warm-up).

• The lesson progresses at a notably slow pace,and students are frequently disengaged or left with nothing meaningful to do.

• Inappropriate or off-task student behavior constantly interrupts or delays the lesson.

T9: Build a Supportive, Learning-Focused Classroom Community

Highly Effective (4)

• Students are invested in the success of their peers. For example, they can be seen collaborating with and helping each other without prompting from the teacher.

• Students may give unsolicited praise or encouragement to their peers for good work, when appropriate.

• Student comments and actions demonstrate that students are excited about their work and understand why it is important.

• There is evidence that the teacher has strong, individualized relationships with students in the class. For example, the teacher might demonstrate personal knowledge of students' lives, interests, and preferences.

• Students may demonstrate frequent positive engagement with their peers. For example, they might show interest in other students' answers or work.

Effective (3)

• Students are invested in their work and value academic success. For example, students work hard, remain focused on learning without frequent reminders, and persevere through challenges.

• The classroom is a safe environment for students to take on challenges and risk failure. For example, students are eager to answer questions, feel comfortable asking the teacher for help, and do not respond negatively when a peer answers a question incorrectly.

• Students are always respectful of the teacher and their peers. For example, students listen and do not interrupt when their peers ask or answer questions.

• The teacher meaningfully reinforces positive behavior and good academic work as appropriate.

• The teacher has a positive rapport with students, as demonstrated by displays of positive affect, evidence of relationship building, and expressions of interest in students' thoughts and opinions.

Minimally Effective (2)

• Students are generally engaged in their work but are not highly invested in it. For example, students might spend significant time off-task or require frequent reminders; students might give up easily; or the teacher might communicate messages about the importance of the work, but there is little evidence that students have internalized them.

• Some students are willing to take academic risks, but others may not be. For example, some students might be reluctant to answer questions or take on challenging assignments; some students might be hesitant to ask the teacher for help even when they need it; or some students might occasionally respond negatively when a peer answers a question incorrectly.

• Students are generally respectful of the teacher and their peers, but there are some exceptions. For example, students might occasionally interrupt, or might be respectful and attentive to the teacher, but not to their peers.

• The teacher may rarely reinforce positive behavior and good academic work, may do so for some students but
not for others, or may not do so in a meaningful way.

• The teacher may have a positive rapport with some students but not others, or may demonstrate little rapport with students

Ineffective (1)

• Students may demonstrate disinterest or lack of investment in their work. For example, students might be unfocused and not working hard, be frequently off-task, or refuse to attempt assignments.

• Students are generally not willing to take on challenges and risk failure. For example, most students might be reluctant to answer questions or take on challenging assignments, most students might be hesitant to ask the teacher for help even when they need it, or students might discourage or interfere with the work of their peers
or criticize students who give incorrect answers.

• Students may frequently be disrespectful to the teacher or their peers. For example, they might frequently interrupt or be clearly inattentive when the teacher or their peers are speaking.

• The teacher may never reinforce positive behavior and good academic work, or s/he may do so for only a few students.

• There may be little or no evidence of a positive rapport between the teacher and the students, or there may be evidence that the teacher has a negative rapport with students.

APPENDIX B

STANDARDS I AND IV OF THE NC TES

Standard I: Teachers Demonstrate Leadership

Observation	Developing	Proficient	Accomplished	Distinguished	Not Demonstrated (Comment Required)
	Element Ia. Teachers lead in their classrooms. Teachers demonstrate leadership by taking responsibility for the progress of all students to ensure that they graduate from high school, are globally competitive for work and postsecondary education, and are prepared for life in the 21st century. Teachers communicate this vision to their students. Using a variety of data sources, they organize, plan, and set goals that meet the needs of the individual student and the class. Teachers use various types of assessment data during the school year to evaluate student progress and to make adjustments to the teaching and learning process. They establish a safe, orderly environment, and they create a culture that empowers students to collaborate and become lifelong learners.				
✓	❑ Understands how they contribute to students graduating from high school. ❑ Uses data to understand the skills and abilities of students.	. . . and ❑ Takes responsibility for the progress of students to ensure that they graduate from high school. ❑ Provides evidence of data driven instruction throughout all classroom activities. ❑ Establishes a safe and orderly classroom.	. . . and ❑ Communicates to students the vision of being prepared for life in the 21st century. ❑ Evaluates student progress using a variety of assessment data. ❑ Creates a classroom culture that empowers students to collaborate.	. . . and ❑ Encourages students to take responsibility for their own learning. ❑ Uses classroom assessment data to inform program planning. ❑ Empowers and encourages students to create and maintain a safe and supportive school and community environment.	
	Element Ib. Teachers demonstrate leadership in the school. Teachers work collaboratively with school personnel to create a professional learning community. They analyze and use local, state, and national data to develop goals and strategies in the school improvement plan that enhances student learning and teacher working conditions. Teachers provide input in determining the school budget and in the selection of professional development that meets the needs of students and their own professional growth. They participate in the hiring process and collaborate with their colleagues to mentor and support teachers to improve the effectiveness of their departments or grade levels.				

	❑ Attends professional learning community meetings. ❑ Displays awareness of the goals of the school improvement plan.	. . . and ❑ Participates in professional learning community. ❑ Participates in developing and/or implementing the school improvement plan.	. . . and ❑ Assumes a leadership role in professional learning community. ❑ Collaborates with school personnel on school improvement activities.	. . . and ❑ Collaborates with colleagues to improve the quality of learning in the school. ❑ Assumes a leadership role in implementing school improvement plan throughout the building.	

Observation	**Element Ic. Teachers lead the teaching profession.** Teachers strive to improve the teaching profession. They contribute to the establishment of positive working conditions in their school. They actively participate in and advocate for decision-making structures in education and government that take advantage of the expertise of teachers. Teachers promote professional growth for all educators and collaborate with their colleagues to				
	Developing	**Proficient**	**Accomplished**	**Distinguished**	**Not Demonstrated (Comment**
	❑ Has knowledge of opportunities and the need for professional growth and begins to establish relationships with colleagues.	. . . and Contributes to the: ❑ improvement of the profession through professional growth. ❑ establishment of positive working relationships.	. . . and ❑ Promotes positive working relationships through professional growth activities and collaboration.	. . . and ❑ Seeks opportunities to lead professional growth activities and decision-making processes.	
	Element Id. Teachers advocate for schools and students. Teachers advocate for positive change in policies and practices affecting student learning. They participate in the implementation of initiatives to improve the education of students.				

	❑ Knows about the policies and practices affecting student learning.	. . . and ❑ Supports positive change in policies and Practices affecting student learning.	. . . and ❑ Participates in developing policies and practices to improve student learning.	. . . and ❑ Actively participates, promotes, and provides strong supporting evidence for implementation of initiatives to improve education.	
	Element Ie. Teachers demonstrate high ethical standards. Teachers demonstrate ethical principles including honesty, integrity, fair treatment, and respect for others. Teachers uphold the *Code of Ethics for North Carolina Educators* (effective June 1, 1997) and the *Standards for Professional Conduct* adopted April 1, 1998.				
	❑ Understands the importance of ethical behavior as outlined in the *Code of Ethics for North Carolina Educators* and the *Standards for Professional Conduct.*	. . . and ❑ Demonstrates ethical behavior through adherence to the *Code of Ethics for North Carolina Educators* and the *Standards for Professional Conduct.*	. . . and ❑ Knows and upholds the *Code of Ethics for North Carolina Educators* and the *Standards for Professional Conduct.*	. . . and ❑ Models the tenets of the *Code of Ethics for North Carolina Educators* and the *Standards for Professional Conduct* and encourages others to do the same.	

Standard IV: Teachers facilitate learning for their students

Observation	Developing	Proficient	Accomplished	Distinguished	Not Demonstrated (Comment
	Element IVa. Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of their students. Teachers know how students think and learn. Teachers understand the influences that affect individual student learning (development, culture, language proficiency, etc.) and differentiate their instruction accordingly. Teachers keep abreast				
		. . . and	. . . and	. . . and	
✓	❑ Understands developmental levels of students and recognizes the need to differentiate instruction.	❑ Understands developmental levels of students and appropriately differentiates instruction.	❑ Identifies appropriate developmental levels of students and consistently and appropriately differentiates instruction.	❑ Encourages and guides colleagues to adapt instruction to align with students' developmental levels.	
✓		❑ Assesses resources needed to address strengths and weaknesses of students.	❑ Reviews and uses alternative resources or adapts existing resources to take advantage of student strengths or address weaknesses.	❑ Stays abreast of current research about student learning and emerging resources and encourages the school to adopt or adapt them for the benefit of all students.	
	Element IVb. Teachers plan instruction appropriate for their students. Teachers collaborate with their colleagues and use a variety of data sources for short- and long-range planning based on the North Carolina Standard Course of Study. These plans reflect an understanding of how students learn. Teachers engage students in the learning process. They understand that instructional plans must be consistently monitored and modified to				
		. . . and	. . . and	. . . and	
✓	❑ Recognizes data sources important to planning instruction.	❑ Uses a variety of data for short- and long-range planning of instruction. Monitors and modifies instructional plans to enhance student learning.	❑ Monitors student performance and responds to individual learning needs in order to engage students in learning.	❑ Monitors student performance and responds to cultural diversity and learning needs through the school improvement process.	
	Element IVc Teachers use a variety of instructional methods. Teachers choose the methods and techniques that are most effective in meeting the needs of their students as they strive to eliminate achievement gaps. Teachers employ a wide range of techniques including information and communication technology, learning styles, and differentiated instruction.				

✓	❑ Demonstrates awareness of the variety of methods and materials necessary to meet the needs of all students.	. . . and ❑ Demonstrates awareness or use of appropriate methods and materials necessary to meet the needs	. . . and ❑ Ensures the success of all students through the selection and utilization of appropriate methods and materials.	. . . and ❑ Stays abreast of emerging research areas and new and innovative materials and incorporates them into lesson plans and	
Observation	**Element IVd. Teachers integrate and utilize technology in their instruction.** Teachers know when and how to use technology to maximize student learning. Teachers help students use technology to learn content, think critically, solve problems, discern reliability, use information, communicate, innovate, and collaborate.				
	Developing	**Proficient**	**Accomplished**	**Distinguished**	**Not Demonstrated (Comment**
✓	❑ Assesses effective types of technology to use for instruction.	. . . and ❑ Demonstrates knowledge of how to utilize technology in instruction.	. . . and ❑ Integrates technology with instruction to maximize student learning.	. . . and ❑ Provides evidence of student engagement in higher level thinking skills through the integration of technology.	
	Element IVe. Teachers help students develop critical-thinking and problem-solving skills. Teachers encourage students to ask questions, think creatively, develop and test innovative ideas, synthesize knowledge, and draw conclusions. They help students exercise and communicate sound reasoning; understand connections;				

✓	❑ Understands the importance of developing students' critical thinking and problem-solving skills.	. . . and ❑ Demonstrates knowledge of processes needed to support students in acquiring critical thinking skills and problem-solving skills.	. . . and Teaches students the processes needed to: ❑ think creatively and critically, ❑ develop and test innovative ideas, ❑ synthesize knowledge, ❑ draw conclusions, ❑ exercise and communicate sound reasoning, ❑ understand connections, ❑ make complex choices, and ❑ frame, analyze and solve problems.	. . . and ❑ Encourages and assists teachers throughout the school to integrate critical thinking and problem solving skills into their instructional practices.	
	Element IVf. Teachers help students work in teams and develop leadership qualities. Teachers teach the importance of cooperation and collaboration. They organize learning teams in order to help students define roles, strengthen social ties, improve communication and collaborative skills, interact with people from different				
✓	❑ Provides opportunities for cooperation, collaboration, and leadership through student learning teams.	. . . and ❑ Organizes student learning teams for the purpose of developing cooperation, collaboration, and student leadership.	. . . and ❑ Encourages students to create and manage learning teams.	. . . and ❑ Fosters the development of student leadership and teamwork skills to be used beyond the classroom.	
	Element IVg. Teachers communicate effectively. Teachers communicate in ways that are clearly understood by their students. They are perceptive listeners and are able to communicate with students in a variety of ways even when language is a barrier. Teachers help students articulate thoughts and ideas clearly and effectively.				
Observation	**Developing**	**Proficient**	**Accomplished**	**Distinguished**	**Not Demonstrated (Comment Required)**

✓ ✓	❑ Demonstrates the ability to effectively communicate with students. ❑ Provides opportunities for students to articulate thoughts and ideas.	. . . and ❑ Uses a variety of methods for communication with all students. ❑ Consistently encourages and supports students to articulate thoughts and ideas clearly and effectively.	. . . and ❑ Creates a variety of methods to communicate with all students. ❑ Establishes classroom practices which encourage all students to develop effective communication skills.	. . .and ❑ Anticipates possible student misunderstandings and proactively develops teaching techniques to mitigate concerns. ❑ Establishes school-wide and grade appropriate vehicles to encourage students throughout the school to develop effective communication skills.	
	Element IVh. Teachers use a variety of methods to assess what each student has learned. Teachers use multiple indicators, including formative and summative assessments, to evaluate student progress and growth as they strive to eliminate achievement gaps. Teachers provide opportunities, methods, feedback, and tools for students to assess themselves and each other. Teachers use 21st century assessment systems to inform instruction and				
✓ ✓	❑ Uses indicators to monitor and evaluate student progress. ❑ Assesses students in the attainment[st] of 21 century knowledge, skills, and dispositions.	. . . and ❑ Uses multiple indicators, both formative and summative, to monitor and evaluate student progress and to inform instruction. ❑ Provides evidence that students attain 21st century knowledge, skills and dispositions.	. . . and ❑ Uses the information gained from the assessment activities to improve teaching practice and student learning. ❑Provides opportunities for students to assess themselves and others.	. . . and ❑ Teaches students and encourages them to use peer and self-assessment feedback to assess their own learning. ❑ Encourages and guides colleagues to assess 21st century skills, knowledge, and dispositions and to use the assessment information to adjust their instructional practice.	